THÈSE

pour l'obtention du Diplôme de

DOCTORAT DE L'UNIVERSITÉ PARIS 7

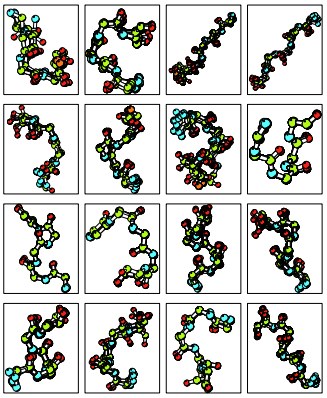
Spécialité : ANALYSES DE GÉNOMES ET MODÉLISATION MOLÉCULAIRE

présentée le 6 février 2001

par

Alexandre de BREVERN

Nouvelles stratégies d'analyse et de prédiction des structures tridimensionnelles des protéines



JURY

Dr Richard LAVERY, CNRS	Président
Pr Gilbert DELEAGE, Université Lyon 1	Rapporteur
Dr Jean-François GIBRAT, INRA	Rapporteur
Pr Serge HAZOUT, Université Paris 7	Directeur de thèse
Dr Jean-Michel CAMADRO, CNRS	Examinateur
Dr Caroline Le VAN KIM, Université Paris 7	Examinateur

Je remercie Richard Lavery d'avoir accepté de présider ce jury.

Je remercie Serge Hazout pour m'avoir donné ce sujet, ainsi que pour la direction et la réalisation de cette thèse, et surtout pour les qualités humaines et la patience dont il a fait preuve tout au long de ce travail.

Je remercie Gilbert Deléage et Jean-François Gibrat pour avoir accepté de juger mon travail et pour leurs remarques très constructives.

Je remercie Jean-Michel Camadro et Caroline Le Van Kim d'avoir accepté d'examiner ce travail.

Je remercie Catherine Etchebest pour son action discrète et constante.

Je remercie l'ensemble de l'Equipe de Bioinformatique Génomique et Moléculaire, avec Pierre "ASSIIRC" Vincens, Pierre "XmMol" Tufféry, France "HXM" Loirat, Anne-Claude "Markov" Camproux-Carat, Fredéric "Ondelettes" Guyon et Joelle "ca compile pas" Hochez pour leurs précieuses collaborations et plus particulièrement Anne Badel-Chagnon pour m'avoir supporté lors de l'ensemble de mes enseignements de statistiques.

Je remercie les stagiaires Romain Gautier, Alexandre Gillet, Emmanuelle Magnifici et Hélène Valadié, qui ont amené leur fraicheur d'esprit dans le laboratoire.

Je remercie l'ensemble de l'équipe des enseignants du DEA d'Analyses de Génomes et Modélisation Moléculaire qui m'ont apportés de précieuses informations et plus particulièrement Philipe Dessen.

Je remercie, bien sur, Marie-Hélène Mucchielli-Giorgi qui m'a permis de mieux comprendre ce qu'est un travail de thèse et de mieux appréhender les problèmes qui en découlent.

Je remercie David Polverari pour les pavés et ses bons mots.

Je remercie Jean-Yves Brossas, Olivier Gourreau, Jacques Tréton et Yves Courtois qui les premiers m'ont fait aimer la recherche.

Je remercie les anciens du DEA qui ont supportés mes questions, Ingrid, Virginie, Emmanuelle, Laurent et Antoine.

Je remercie Nicolas, Fred, Pierre, Franck et Jérome pour avoir écouté "avec bonté" mes divagations alors qu'ils n'y ont jamais compris grand chose.

Enfin, je remercie mes parents pour leur soutien indéfectible et leur aide.

Sommaire

Chapitre 1

Introduction

En 1995, j'effectuais un stage de Biologie Moléculaire dans l'unité du Dr. Jean-Yves Courtois. Avec Jean-Yves Brossas et le Dr. Jacques Tréton, nous tentions alors de caractériser une recircularisation de l'ADN mitochondrial. Pour ce faire, nous séquençions alors quelques centaines de bases "à la main avec du phosphore radioactif". Actuellement, plus d'une trentaine de génomes complets sont disponibles. Les progrès technologiques et méthodologiques nous ont donné une masse d'informations extraordinaire qu'il faut désormais traiter.

Avec cette explosion récente des bases de données aussi bien génomiques que protéiques, se pose de manière cruciale le problème de la *génomique structurale*, du repliement des protéines. Les protéines sont formées par la succession d'acides aminés liés entre eux par des liaisons peptidiques. Cette succession est directement liée à la séquence génomique et contrôle la formation de la structure tridimensionnelle de la protéine. Cette structure contrôle l'ensemble des fonctions biologiques des protéines. La connaissance et la possibilité de prédire leurs structures sont donc fondamentale.

Possédant la séquence d'une protéine, nous devrions pouvoir en déduire directement sa structure tridimensionnelle. Les facteurs physico-chimiques et cinétiques qui agissent lors du repliement d'une protéine sont toutefois encore trop complexes et les approximations trop grandes pour pouvoir simuler un repliement *in silico* à l'identique de celui *in vivo* [119].

Aussi, faut-il obligatoirement connaître une protéine possédant une certaine homologie avec celle qui nous intéresse pour pouvoir travailler. La structure tridimensionnelle des protéines peut être caractérisée en ne tenant compte que de ses parties régulières et répétitives, les feuillets et les hélices. Elles décrivent une géométrie spatiale stabilisée par des liaisons internes. Le reste, plus variable, est dénommé boucles. Les méthodes de prédiction des ces trois états ont

connu ces dernières années grâce à l'utilisation de séquences proches et de réseaux neuronaux une augmentation importante de leur taux de prédiction permettant d'atteindre actuellement environ 75 % de prédiction correcte [166, 168, 26].

Un alphabet à 3 états demeure cependant assez pauvre structurellement. Différentes équipes ont ainsi décidé d'approfondir la spécificité structurale, en analysant les connections qui existent entre deux structures régulières consécutives le long des protéines. Des librairies de boucles plus ou moins longues ont donc été proposées [203, 114, 13, 204]. La composition très spécifiques de certaines boucles ont permis des prédictions intéressantes [189]. Toutefois, ces différentes méthodes sont liées à la définition des structures secondaires.

Pour décrire plus précisément la structure des protéines, différentes équipes ont élaboré des alphabets structuraux comportant un nombre différent de prototypes. Unger et collaborateurs [198] et Schuchhardt et coll. [175] en proposent une centaine, ceux-ci permettant une bonne précision dans l'approximation de la structure, difficilement utilisable néanmoins dans une approche de prédiction. Rooman et coll. [162], Fetrow et coll. [53], Bystroff et Baker [19], et Camproux et coll. [23, 24] en proposent un nombre plus restreint allant de 4 à 13 prototypes. Tous montrent les spécificités en acides aminés selon le prototype observé. Bystroff et Baker [19] propose en plus une méthode de prédiction.

Dans le cadre de cette thèse, nous avons donc décidé de développer une nouvelle méthodologie pour concevoir des séries de blocs protéiques dans un but tant de prédiction locale que d'approximation des structures protéiques. Ayant obtenu différentes séries de prototypes, la série qui sera choisie devra permettre une prédiction avec un taux acceptable et aussi permettre une reproduction correcte de la structure protéique. Il conviendra alors de vérifier et d'analyser la spécificité des blocs protéiques au le plan structural, voir leur stabilité et les comparer aux différents alphabets existants.

Notre second objectif est d'utiliser cet "alphabet structural" (ensemble des blocs protéiques) dans une prédiction locale du squelette polypeptidique. Cette prédiction sera effectuée avec une méthode qui permet de comprendre l'importance des acides aminés de manière simple. Pour améliorer cette prédiction, nous nous sommes basés sur deux concepts : (i) *1 repliement local* $\rightarrow$ *n séquences* et (ii) *1 séquence* $\rightarrow$ *n repliements*. Le premier concept signifie que plusieurs types de séquences peuvent être associés à la même structure et le second qu'une séquence peut être associée à plusieurs type de repliements [186]. Ces deux aspects seront développés en se

basant sur la recherche d'un indice de fiabilité lié à la prédiction locale, pour trouver des zones de fortes probabilités.

Ayant défini un alphabet structural, certains mots, i.e. successions de blocs protéiques peuvent apparaître plus fréquemment que d'autres ainsi que l'avaient remarqué Fetrow et collaborateurs [53]. Nous essayerons alors de définir au mieux quelle est l'architecture de ces successions, les liens existants entre ces différents mots.

Les étapes précédentes avaient pour but de caractériser un alphabet structural permettant de décrire au mieux la structure 3D des protéines et les dépendances entre blocs protéiques. Du fait de cette redondance qui peut apparaître dans la structure protéique, il nous a semblé intéressant de mettre au point une méthode de compactage qui permet d'associer des structures proches sur le plan tridimensionnel local. Cette approche appelée "protéine hybride" de conception simple permet de catégoriser en classes "structuralement dépendantes". Cette approche, en plus du compactage, peut être utilisée dans une optique différente, celle de la recherche d'homologie structurale et de la caractérisation des dépendances entre structures et séquences.

Le travail réalisé tourne ainsi autour de la définition d'un alphabet structural, et, de ses potentiels dans le cadre de la prédiction de structure à un niveau local, de la recherche d'homologie structurale et de la caractérisation des dépendances entre structures et séquences. Les différents chapitres détailleront ces différents aspects.

Chapitre 2
Les Protéines

Les protéines représentent plus de la moitié du poids sec des cellules. Leur importance biologique vient de leur aptitude à reconnaître d'autres molécules avec lesquelles elles s'associent de manière transitoire d'après leur configuration spatiale. Ce premier chapitre récapitulera les principales bases biochimiques des protéines, et s'attachera à différentes méthodes d'analyse et de prédiction de leur structure tridimensionnelle. La structure d'une protéine est essentiellement déterminée par sa séquence primaire en acides aminés. Dans une cellule, une séquence protéique donne lieu à un seul type de structure tridimensionnelle. Toutefois, il est particulièrement difficile encore à l'heure actuelle de prévoir la structure d'une protéine en se basant uniquement sur sa séquence [119].

2.1 La structure protéique

2.1.1 Les acides aminés

Les protéines sont des polymères, des macromolécules biologiques primordiales dans l'ensemble du règne animal et végétal. Depuis la découverte de la structure de l'ADN (Acide Désoxyribo Nucléique), le dogme de la biologie moléculaire peut se résumer en trois points:

(1) l'ADN porte l'information génétique,

(2) cette information est transcrite en ARN mono-brin qui possède une information souvent codante,

(3) avec l'aide du complexe ribosomique, l'ARN est traduit en protéine.

Les protéines sont des successions d'acides aminés simples. Il en existe 20 principaux. Un

acide aminé est composé d'un carbone asymétrique dit carbone α (C_α). Ce carbone tétravalent est lié (cf. figure 2.1) à une fonction amine NH_2 (*N: azote et H: hydrogène*), à un atome d'hydrogène H, à une fonction acide COOH et à un radical, un groupement de taille plus ou moins important appelé chaîne latérale (R).

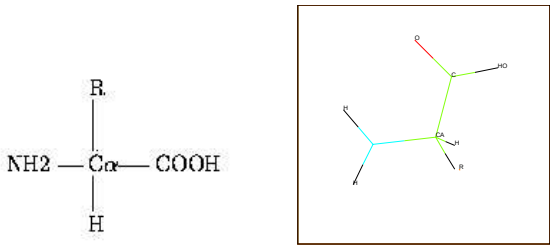


FIG. 2.1 – *Structure schématique d'un acide aminé à gauche plane et à droite visualisé en 3D par le logiciel XmMol [194]*

Les acides aminés libres sont solubles dans l'eau du fait des groupements polaires qu'ils portent. Il existe plusieurs types de chaînes latérales (R) qui se différencient selon leurs propriétés physico-chimiques :

- R hydrophobes : Alanine, Valine, Leucine, Isoleucine, Phénylalanine, Proline et Méthionine.
- R chargés : Aspartate, Glutamate, Lysine et Arginine.
- R polaires : Sérine, Thréonine, Cystéine, Acide Aspartique, Acide Glutamique, Histidine, Tyrosine et Tryptophane.
- sans R : Glycine.

Ils se différencient aussi suivant leur encombrement stérique (cf. Tableau 2.1). Certains possèdent des cycles aromatiques comme la Phénylalanine, la Tyrosine et le Tryptophane. La Proline est le seul acide aminé dont la chaîne latérale est liée aussi à l'azote du squelette polypeptidique et donc dépourvu d'hydrogène. La Cystéine possède un groupement souffré qui peut s'oxyder, ce qui permet la création de pont disulfure qui peut stabiliser de manière primordiale la structure protéique. De nombreuses études portant sur les relations et les équivalences au niveau physico-chimique des acides aminés ont été menées. On peut noter le travail de William

Taylor qui a mis au point une méthode de représentation simple prenant en compte à la fois les propriétés physico-chimiques, mais aussi le volume des acides aminés (cf. figure 2.2 [190]).

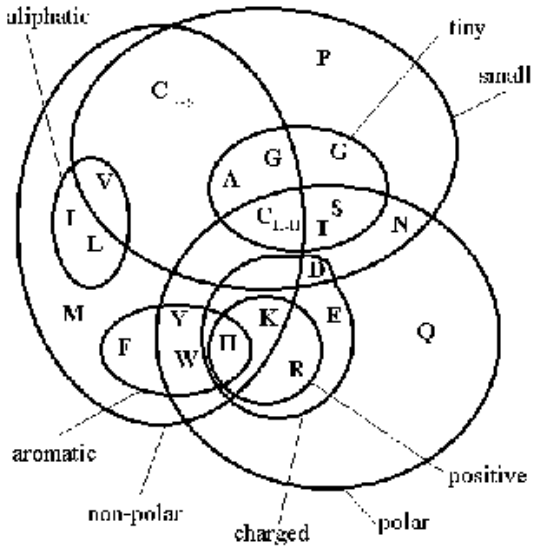


FIG. 2.2 – *Diagramme de Venn portant sur l'équivalence entre les différents acides aminés par rapport à leurs propriétés physico-chimiques et leurs volumes*

Deux acides aminés vont se lier en créant une liaison polypeptidique (cf. Figure 2.3), par condensation entre le groupement carboxyle du premier et le groupement amine du second. C'est une polymérisation qui s'effectue par perte d'une molécule d'eau. Cette nouvelle liaison implique une rigidité importante. Quand le nombre d'acides aminés est faible, la macromolécule est souvent nommée oligopeptide. Un nombre important de résidus est nécessaire pour parler de polypeptide ou protéine (les chiffres varient dans la littérature entre 20 et 50 résidus).

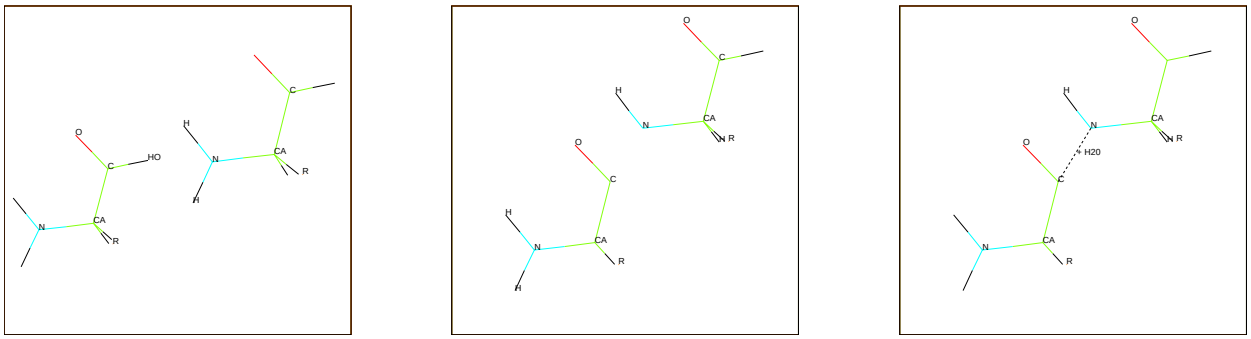


FIG. 2.3 – *Polymérisation de deux acides aminés par condensation et création d'une liaison peptidique.*

Nom	Name	3	1	masse	surface	volume	form
Alanine	Alanine	ALA	A	71.09	115	88.6	CH_3
Arginine	Arginine	ARG	R	156.19	225	173.4	HN=
Acide Aspartique	Aspartic Acid	ASP	D	114.11	150	111.1	HO(
Asparagine	Asparagine	ASN	N	115.09	160	114.1	H_2N
Cystéine	Cysteine	CYS	C	103.15	135	108.5	HS-(
Acide Glutamique	Glutamic Acid	GLU	E	129.12	190	138.4	HO(
Glutamine	Glutamine	GLN	Q	128.14	180	143.8	H_2N
Glycine	Glycine	GLY	G	57.05	75	60.1	H_2N
Histidine	Histidine	HIS	H	137.14	195	153.2	N*H

Nom	Name	3	1	masse	surface	volume	form
Isoleucine	Isoleucine	ILE	I	113.16	175	166.7	CH_3
Leucine	Leucine	LEU	L	113.16	170	166.7	$(CH_2$
Lysine	Lysine	LYS	K	128.17	200	168.6	H_2N
Méthionine	Methionine	MET	M	131.19	185	162.9	CH_3
Phénylalanine	Phenylalanine	PHE	F	147.18	210	189.9	Aro-
Proline	Proline	PRO	P	97.12	145	112.7	N*H
Sérine	Serine	SER	S	87.08	115	89.0	HO-
Thréonine	Threonine	THR	T	101.11	140	116.1	CH_3
Tryptophane	Tryptophane	TRP	W	186.12	255	227.8	Aro*
Tyrosine	Tyrosine	TYR	Y	163.18	230	193.6	HO-
Valine	Valine	VAL	V	99.14	155	140.0	$(CH_2$

TAB. 2.1 – *les 20 acides aminés classiques avec leur nom, en français (Nom), nom des acides aminés (3),le code à une lettre (1), leur masse, leur surface, u fermeture du cycle et Aro pour un groupement aromatique) et enfin une représe*

On distingue alors couramment la région axiale, monotone et de composition constante (carbones et azote), appelée squelette de la chaîne polypeptidique et la succession des chaînes latérales variables selon le type d'acides aminés.

L'extrémité N-terminale correspondant au premier acide aminé de la protéine porte un groupement amine encore libre et l'extrémité C-terminale le dernier acide aminé de la protéine avec un groupement carboxyle libre aussi. Ces deux groupements à pH physiologique sont le plus souvent ionisés. La synthèse s'effectue de l'extrémité N-terminale vers l'extrémité C-terminale.

2.1.2 Les différents niveaux de la structure protéique

Les protéines peuvent-être subdivisées en trois catégories principales: (i) des protéines solubles en général globulaires, le plus souvent compartimentales (cytoplasme, noyau, mitochondrie et chloroplaste), (ii) des protéines membranaires qui sont incluses dans les membranes lipidiques, et (iii) des protéines fibreuses nécessaires pour des actions biologiques particulières telles la contraction musculaire ou le maintien du cytosquelette. Parmi les protéines membranaires, certaines possèdent à la fois des segments totalement enfouis et des régions qui baignent dans le cytoplasme.

Le calcul systématique des angles et des distances de tous les atomes des protéines cristallisées montrent qu'un certain nombre d'angles et de distances restent assez constant (cf. figure 2.4a), alors que d'autres sont beaucoup plus variables. Il s'agit par exemple des angles de valence et des longueurs de liaison tandis que les angles dièdres (ou de torsion) assurent la diversité du repliement 3D et sont donc nettement plus variables.

Différents angles sont utilisés pour décrire les protéines :

- les angles dièdres : ϕ, ψ et ω, qui décrivent le squelette; en prenant en compte pour l'angle ϕ_x (ϕ en position x dans la séquence), les atomes C'$_{x-1}$, N$_x$, Cα_x et C'$_x$, pour ψ_x,les atomes N$_x$, Cα_x et C'$_x$ et N$_{x+1}$ et pour ω, Cα_x, C'$_x$, N$_{x+1}$ et Cα_{x+1}. La figure 2.4c montre leur positionnement sur la chaîne polypeptidique.

- l'angle α : angle formé par 4 carbones α successifs.

- l'angle τ : angle formé par 3 carbones α successifs.

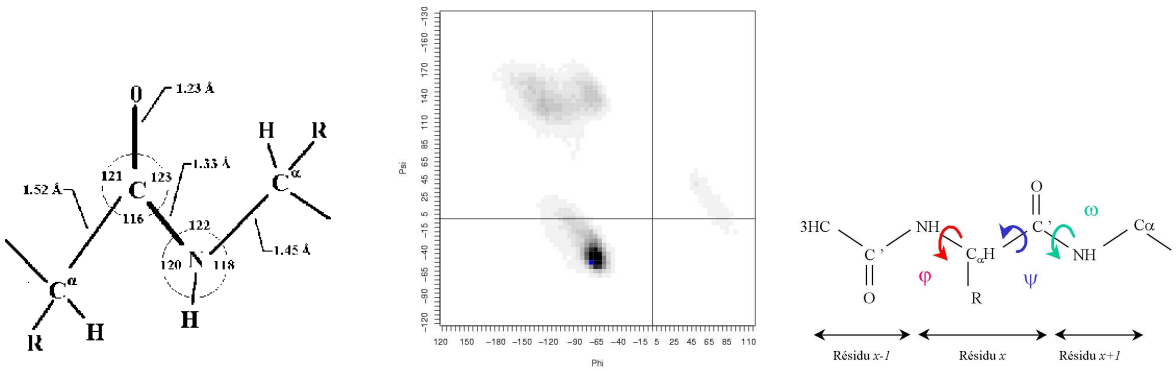


FIG. 2.4 – *(a) Valeurs classiques de certains angles et distances restant constantes dans les protéines. (b) Diagramme de Ramachandran, avec l'angle ϕ en abcisse et l'angle ψ en ordonnée. (c) Schéma des angles ϕ, ψ et ω.*

Les angles ϕ et ψ sont classiquement représentés l'un par rapport à l'autre dans un diagramme de Ramachandran (cf. figure 2.4b), qui fut utilisé pour la première fois en 1968 par G.N. Ramachandran se servant à l'époque de modèles énergétiques pour caractériser les zones préférentielles de ces angles de torsions. La succession des acides aminés correspond à la séquence primaire (1D). La structure tridimensionnelle résultant du repliement de la protéine est appelée structure tertiaire (3D). Nous nous attarderons sur une classification intermédiaire notée structure secondaire (2D). La base de données qui contient la totalité des structures 3D protéiques se nomme la Protein Data Bank ou PDB [6].

2.1.2.1 La structure secondaire

La structure secondaire se caractérise classiquement par les hélices α et les feuillets β, structures possédant des caractéristiques répétitives d'intérêt.

hélices et feuillets Hélice α et feuillet β sont les deux formes répétitives les plus importantes des protéines représentant chacune respectivement 30 et 20 % des résidus. Cette importance vient de leur stabilité énergétique particulière [184].

Les hélices α tournent dans le sens dit "main droite". Les chaînes latérales sont situées à l'extérieur de l'hélice. L'ensemble s'inscrit dans un cylindre de 10,5 Å de diamètre, le tour de spire ou pas de l'hélice fait 5,4 Å soit 3.6 résidus. L'hélice α est une structure thermodynamiquement stable du fait de nombreuses liaisons hydrogènes entre groupements amines (NH_2) et

carboxyles($C = O$). La figure 2.5a montre la structure de l'hémoglobine (code PDB : 1bbb) qui est une protéine pratiquement tout-α, ces hélices assurent intégralement la fonction biologique de la protéine. La figure 2.5b montre le domaine 3 de l'hélicase de *Escherichia coli* (code PDB: 1cuk) qui est un domaine comportant trois hélices α. La figure 2.5c montre le squelette de la chaîne polypeptidique. La répétitivité du motif est nette, les groupements carboxyles sont presque parallèles. Les liaisons hydrogènes existantes entre les résidus sont indiquées en pointillés. Ce sont des liaisons entre l'oxygène d'un résidu en position i et le groupement azoté d'un résidu en position $i+4$; cette liaison est notée [i:i+4].

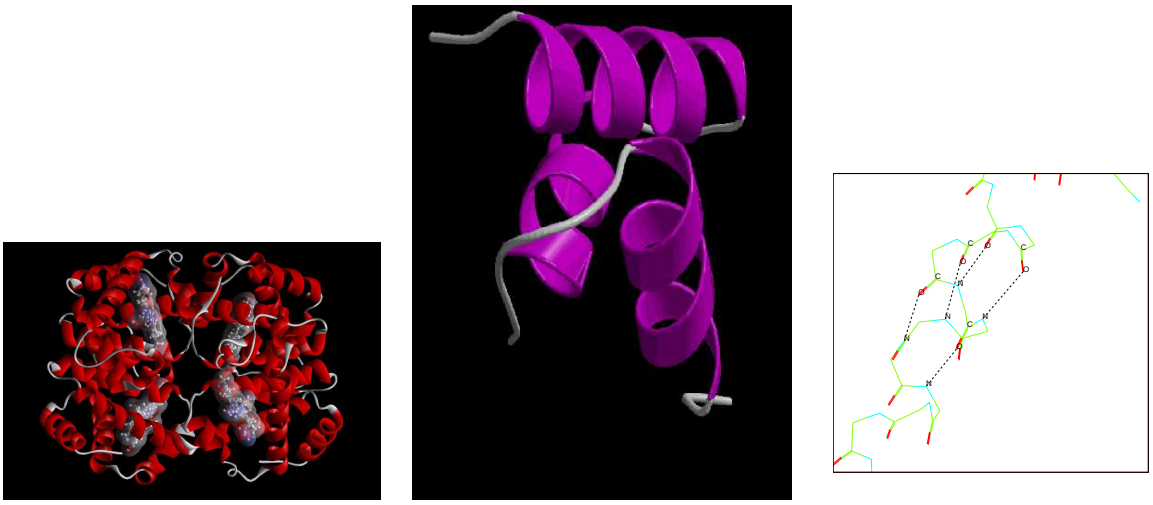


FIG. 2.5 – *a. Une proteine composée en majorité d'hélice α, la 1bbb (les hélices sont en rouges et maintiennent les hèmes). b. Squelette de la première hélice d'une hélicase de Escherichia coli (code PDB: 1cuk). c. Les liaisons entre les résidus 154 à 163 de cette protéine sont indiquées en pointillés.*

Les hélices, malgré leur définition, ne sont pas des tubes rigides. Plus d'un quart des hélices α sont fortement non-régulières [3]. Un lien direct entre la longueur de l'hélice et son degré de courbure a d'ailleurs été déterminé [113].

Les hélices possèdent une composition en acides aminés particulière. De nombreuses études ont montré aussi l'existence d'acides aminés sur-représentés aux extrémités C- et N-terminales des hélices. Une dizaine de successions spécifiques qui forment des terminaisons stables et qui induisent ces fins d'hélices ont été caractérisées [154, 2]. En plus des acides aminés classiques du type Proline ou Glycine, certains acides aminés ont un rôle structural dans ces motifs. Les Glycines se trouvent dans la zone des Ψ positifs (cf. figure 2.4b). Les hélices sont présentes

dans la zone des Φ et Ψ négatifs. Certaines classes d'hélices α ont été étudiées du fait de leur importance fonctionnelle comme les hélices amphipatiques (ayant une face polaire et une autre non-polaire) [176].

Les feuillets sont moins stables thermodynamiquement que les hélices α [184]. Ils sont dus à un aller-retour de la chaîne polypeptidique qui fait que les segments deviennent adjacents. Ils se retrouvent dans la zone des Φ négatifs et Ψ positifs (cf. figure 2.4b).

La figure 2.6a montre une porine, protéine tout-β transmembranaire, son ouverture permet le passage de divers solutés. La figure 2.6b représente le premier domaine de la protéine ldu TNF (code PDB : 1ext) d'*Escherichia coli*, où se trouvent plusieurs feuillets β. La figure 2.6c montre un agrandissement du feuillet comprenant les résidus 126-130 et 150-154. Les liaisons stabilisantes existant entre les groupements CO et NH ont été notées en pointillés. Quand les segments sont orientés dans des directions opposées, les feuillets sont dits anti-parallèles. S'ils sont orientés dans la même direction, ils sont dits parallèles.

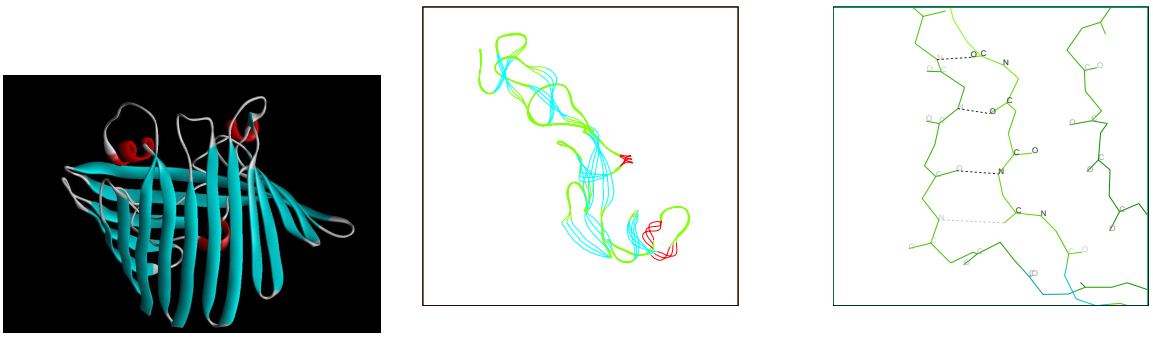


FIG. 2.6 – *(a) Représentation d'une porine, protéine fortement β, les feuillets sont en bleu. (b) Domaine 1 extracellulaire du récepteur au TNF (code PDB: 1ext) fortement β. (c) Liaisons stabilisant un feuillet β.*

Ces deux conformations ne font pas intervenir les chaînes latérales mais seulement le squelette polypetidique. Ceci distingue la structure secondaire de la structure tertiaire. Les hélices α représentent environ 30% des protéines, les feuillets β environ 20%. Le reste est souvent dénommé boucles et a fréquemment été considéré comme variable. Toutefois, on trouve d'autres formes régulières.

structures	nombre de résidus par tours	tours par résidu (Å)	rayon(Å)	Φ	Ψ
hélices α	3.6	1.5	2.3	-57	-47
feuillets β anti-parallèle	2.0	3.4	1.0	-139	135
feuillets βparallèle	2.0	3.2	1.0	-119	113

TAB. 2.2 – *valeurs classiques des angles des hélices α et des feuillets β*

L'attribution des structures secondaires est souvent un problème délicat. Cinq méthodes d'assignation sont actuellement utilisées :

(a) DSSP[100] (1983): Cette méthode est fondée sur une recherche des liaisons hydrogènes spécifiques des structures répétitives. Pour les hélices α, DSSP observe les liaisons en positions $(i, i+3)$ ou $(i, i+4)$, alors que pour les feuillets β des liaisons plus éloignées sont en jeu. DSSP ne prend en compte que des structures secondaires répétitives d'au moins 4 résidus de long.

(b) DEFINE[156] (1988): Elle se base sur les distances entre carbones α. Ces distances sont comparées à des distances de structures secondaires connues. Quand au moins 4 distances successives correspondent à un même type de structure secondaire, la structure secondaire est attribuée au fragment. Les auteurs décrivent des "super" structures secondaires, qui ne sont pas implémentées dans le logiciel disponible.

(c) P-CURVE[181] (1989): Les protéines sont décrites par un axe global, qui suit au mieux le squelette de la protéine. Cet axe est calculé avec une fonction de minimisation qui prend en compte la position des atomes de la chaîne polypeptidique et qui passe au mieux entre eux. Les structures secondaires sont alors définies en comparant les données obtenues avec des segments de structures idéales pour chaque structure secondaire.

(d) Consensus[31] (1993): Les trois méthodes précédentes se basent sur des critères distincts. Colloc'h et collaborateurs ont observé que les assignations qu'ils donnaient étaient parfois fort divergentes. Plus d'un tiers des résidus ne sont pas assignés de la même façon par les trois algorithmes. Ceci peut poser des problèmes, principalement pour la prédiction de ces structures. Les différences se situent principalement aux extrémités des structures répétitives qui sont traitées de manière différente selon les algorithmes. Les assignations incompatibles ("α" pour un algorithme "β" pour un autre) sont en nombre négligeable. Le consensus revient à choisir dans les cas litigieux la solution donnée par deux algorithmes.

	DSSP (%)	P-CURVE (%)	DEFINE (%)	STRIDE (%)	concensus (%)
hélice	93,8	92,4	83,6	93,2	94,2
feuillet	78,4	74,4	66,8	77,5	79,4
boucles	79,3	80,6	84,4	81,2	85,1
Total	83,4	82,4	79,0	84,1	86,5

TAB. 2.3 – *Assignement comparé entre DSSP, P-CURVE, DEFINE, STRIDE et la méthode consensus pour chacun des trois types de structures secondaires avec P-SEA (Table II, page 293 [116]).*

(e) STRIDE[58] (1995): STRIDE ajoute un critère angulaire défini dans DSSP [100] au notion des liaisons hydrogènes qui sont normalement présentes dans les structures répétitives. Pour cela, les mesures adéquates sont directement reliées à celles utilisées dans des travaux cristallographiques. En plus des hélices α et des feuillets β, ils utilisent leur approche pour définir d'autres types de structures caractéristiques. Celles-ci seront développées dans le paragraphe suivant.

(f) P-SEA[116] (1997): L'assignation des structures secondaires par P-SEA (pour *Protein Secondary Element Assignement*) s'effectue exclusivement sur les positions des carbones α. Trois distances et deux angles sont calculés avec d2i pour la distance Cα_{i-1} Cα_{i+1}, d3i pour la distance Cα_{i-1} Cα_{i+2}, d4i pour la distance Cα_{i-1} Cα_{i+3}, les angles α_i décrivant les carbones *i-1, i, i+1, i+2* et τ_i décrivant les carbones *i-1, i, i+1*. L'assignation se fait si les distances et/ou les angles correspondent à des valeurs types décrivant les hélices et les feuillets.

Comparaison entre les différents algorithmes :

L'ensemble de ces méthodes, à l'exception du consensus, dépend d'un certain nombre de définitions sur la valeur des angles ou des distances et des variations autorisées autour de ces valeurs. Les différentes méthodes d'assignation sont donc loin d'être équivalentes. La méthode consensus dérive d'ailleurs de cette constatation, seulement 64 % des résidus assignés par les trois méthodes DSSP, P-CURVE et DEFINE étant attribués au même type de structure secondaire, et propose de prendre en compte l'ensemble des 3 algorithmes précédents.

Le tableau 2.3 récapitule les attributions comparées suivant le type de structure assigné par P-SEA. De fortes différences sont visibles principalement pour les feuillets β qui sont les plus difficiles à caractériser. Dans le tableau 2.4, j'ai récapitulé les concordances d'assignation entre les 5 algorithmes actuellement disponibles (PSEA, DSSP, STRIDE, DEFINE, PCURVE)

	PSEA	DSSP	STRIDE	DEFINE	PCURVE
PSEA	—	74,81	83,03	68,73	76,81
DSSP		—	87,25	63,30	67,53
STRIDE			—	66,43	73,53
DEFINE				—	62,32

TAB. 2.4 – *Fréquence d'assignation commune (en %) entre différents algorithmes d'assignation de structures secondaires effectuée avec 906 chaînes issues de la base de données de R. Dunbrack possédant moins de 50% de similitudes de séquences. Les différentes versions des logiciels utilisés sont mises entre parenthèse avec P-SEA (version 2.0), P-CURVE (version 3.1), DSSP (version DsspCMBI-April-2000, modification du programme original en 1988 et 1994), STRIDE (version 1995) et DEFINE (version 2, 1994).*

type de structures	Φ	Ψ	sur les résidus	liaison CO-NH
hélices 3_{10} *classique*	-71	-18	*i+1* et *i+2*	*i* et *i+3*
hélices 3_{10} *"parfaite"*	-74	-4	*i+1* et *i+2*	*i* et *i+3*
hélices π	-57	-70	*i+1*, *i+2* et *i+3*	*i* et *i+5*

TAB. 2.5 – *valeurs nécessaires à l'assignation des hélices 3_{10} et π.*

pour une base de données de 906 chaînes protéiques possédant moins de 50% de similitudes de séquences. Les fichiers ne pouvant être analysés par un programme donné n'ont pas été pris en compte. Les résultats obtenus sont en forte concordance avec ceux de la littérature. Un maximum de similitude se trouve entre l'assignation effectuée par DSSP et STRIDE, ce qui est logique du fait de l'utilisation par STRIDE des valeurs angulaires définies dans DSSP. Le logiciel qui possède une assignation la plus divergente des autres méthodes est le logiciel DEFINE qui utilise uniquement des distances sur les C_α. Sa notice d'utilisation dit d'ailleurs qu'il est loin d'avoir les critères optimaux nécessaires à l'assignation. Par ailleurs, il a le taux le plus élevé d'impossibilité de lecture de fichiers PDB.

Les autres formes régulières Les hélices 3_{10} et π sont deux formes largement moins présentes que les hélice α dans les protéines. Le tableau 2.5 récapitule les valeurs classiques de ces deux conformations. Les hélice 3_{10} (cf. figures 2.7a et 2.7b) possèdent un diamètre moins important que les hélices α, et sont par ailleurs souvent incluses dans l'une d'elles. Elles représentent entre 3 et 4 % des résidus des structures protéiques, ce qui en fait la quatrième structure secondaire la plus importante.

Des approches théoriques tendent à montrer que leur taille est limitée à quelques résidus du fait de contraintes énergétiques [157]. Un des facteurs importants dans la création lors

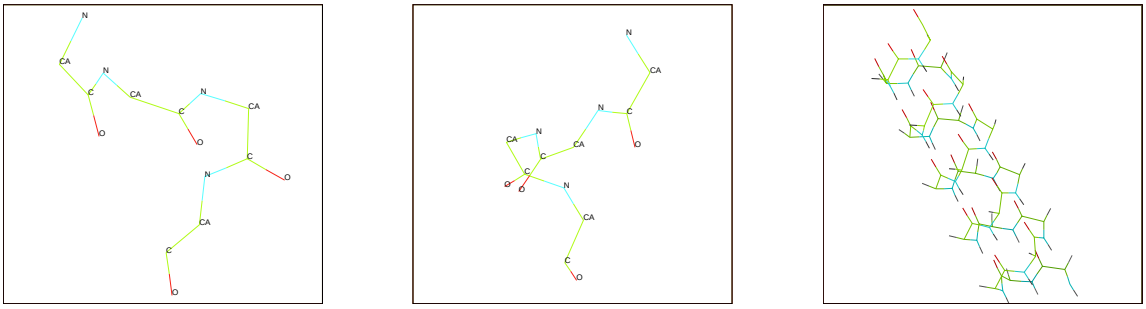


FIG. 2.7 – *Représentations (a) d'une hélice 3_{10} ayant des caractéristiques classiques (AZW chaîne A résidus 51 à 54), (b) une hélice 3_{10} ayant des caractéristiques "parfaites" (protéine 1VPN chaîne C résidus 134 à 137) et (c) une hélice π (modèle théorique construit par le Pr. Etchebest) visualisées avec XmMol [194].*

du repliement d'une hélice 3_{10} est l'absence de liaisons entre les résidus i et $i+4$ [188]. Leur diamètre diffère des hélices α. Leur composition en acides aminés induit des extrémités C- et N-terminales plus étendues [103] que pour les hélices α. Deux définitions existent des hélices 3_{10} (cf. tableau 2.5). La première correspond à la définition classique, moyenne qui est observée dans les protéines. La seconde est dite "parfaite" et représente ce qui serait énergétiquement le plus favorable. Ce second type d'hélice ne correspond qu'à 1 - 2 % des hélices 3_{10} présentes dans la PDB d'après une recherche exhaustive que j'ai menée sur la base de données des structures protéiques (PDB) de juin 2000.

Tout comme les hélices 3_{10}, les hélices π (cf. figures 2.7c) sont fortement impliquées dans les transitions des hélices α vers les boucles. Ces transitions sont de plusieurs types. Les plus importantes sont celles qui partent vers la zone α_L (zone dites des "Glycines" avec un angle dièdre ϕ positif), ensuite ce sont celles qui partent vers la zone α_R (zone classique des hélices α) [150].

Les boucles et les connections entre structures secondaires En dehors des structures répétitives, d'autres structures ont été découvertes, comme les coudes dont le nombre de catégories est d'une huitaine [202]. Les plus classiques sont présentées sur la figure 2.8. Les angles et liaisons les caractérisant sont reportés dans le tableau 2.6. Ils sont retrouvés plus ou moins couramment selon le type de coudes, certains très rares, d'autres spécifiques de fin de protéines [51] ou de conformation spécifique [5]. Ces régions ont toujours toutefois une taille limitée. Le logiciel STRIDE permet de déterminer l'attribution d'un certain type de coudes.

type de structures	Φ	Ψ	sur les résidus	liaison CO-NH
coudes de type I	-60	-30	*i+1*	*i* et *i+3*
	-90	0	*i+2*	
coudes de type II	-60	+120	*i+1*	*i* et *i+3*
	-80	0	*i+2*	
coudes de type III	-60	-30	*i+1*	*i* et *i+3*
	-80	-30	*i+2*	
coudes gamma	+70	-60	*i+1*	*i* et *i+2*

TAB. 2.6 – *valeurs nécessaires à l'assignation des coudes de types I à III et gamma.*

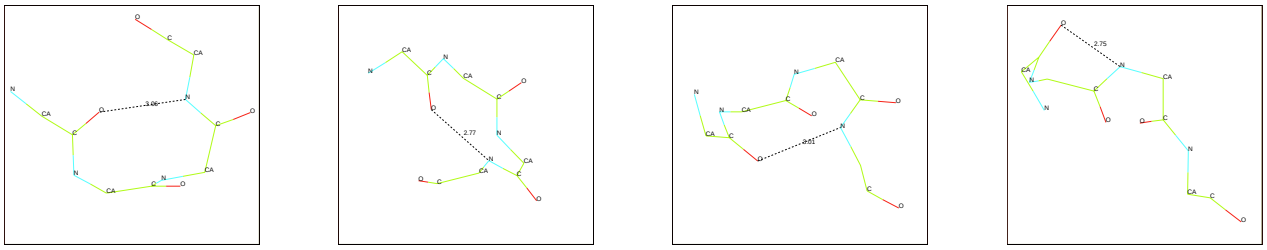


FIG. 2.8 – *Représentations (a) d'un coude de type I (protéine 125L chaîne A résidu 121à 124), (b) d'un coude de type II (1ALL chaîne A résidu 112 à 115), (c) d'un coude de type III (109M chaîne A résidu 13 à 16) et (d) d'un coude de type gamma (1SCY chaîne A résidu 1 à 4).*

Des classifications des boucles donnent lieu à de nombreuses et différentes taxonomies dont des termes identiques recoupent parfois des réalités distinctes [46, 51].

Des classifications des boucles entre structures secondaires α et/ou β ont permis d'observer des spécificités en acides aminés pour certaines boucles [203, 114, 13] et l'établissement de procédures de prédiction [204]. Toutefois, l'ensemble de ces méthodes est basé sur la définition de points de début et de fin des structures secondaires qui sont parfois sujettes à caution et ont des difficultés à considérer l'ensemble des boucles, se limitant à des tailles inférieures à 9 résidus [44, 63].

2.1.2.2 Le nombre de repliements

L'utilisation des structures secondaires permet de catégoriser les protéines en différentes familles, tout-α, tout-β, α/β, $\alpha+\beta$, non classifiable [127]. Elle a permis ainsi de créer de véritables arbres reliant des sous-familles structurales, donnant lieu à des classifications parfois totalement supervisées comme pour SCOP [132] ou automatisées en grande partie comme CATH [136] ou FSSP [86]. L'intérêt de ces approches n'est pas une simple classification mais permet aussi de

trouver de nouveaux liens entre structure, séquence et fonctions [131, 87, 89, 73, 139, 140, 123], et ainsi d'estimer la variété de repliements qui existent dans les protéines [138].

Des études plus théoriques basées sur des méthodes statistiques montrent qu'un certain pourcentage de repliements (70 %) possède une spécificité séquentielle suffisante pour impliquer un repliement local particulier [177]. Le reste est trop peu déterminé et ne pourrait donc pas servir dans le cas d'une prédiction de la structure 3D protéique à partir de la séquence.

Le nombre global de repliements est estimé à un peu plus de 2000 [69]. Toutefois, seuls 900 seraient vraiment différents. Plus récemment, les mêmes auteurs augmentent leurs chiffres à 4000 repliements. 375 de ces repliements représenteraient 70 % des repliements existants dans les protéines, les plus courants, il en faudrait au final 930 pour en représenter 90% [70].

2.2 Prédictions de la structure protéique

2.2.1 La structure secondaire

Deux grandes familles de prédiction de la structure secondaire existent avec les méthodes statistiques, les plus anciennes, et les méthodes "d'apprentissage", les réseaux de neurones. Une nouvelle catégorie est apparue plus récemment, les méthodes consensus qui prennent des résultats de plusieurs algorithmes différents. Plus de milles articles sur la prédiction des structures secondaires ayant été publiés à ce jour, les quelques exemples qui suivent ont donc un caractère arbitraire mais sont, me semble-t-il, les plus représentatifs.

2.2.1.1 Méthodes statistiques

Les méthodes statistiques ont été les premières mises en place. Elles se basent sur une utilisation directe de la composition en acides aminés et se sont complexifiées avec l'utilisation de la théorie de l'information et des alignements de séquences.

- Chou et Fasman [29] : La première méthode de prédiction se base sur la tendance de chaque acide aminé à se trouver associé à une structure hélice α, brin β ou boucle. Cette méthode se base sur le calcul d'un coefficient de corrélation pour chaque type d'acide aminé dans chaque type de structures secondaires. Le taux de prédiction réel de la méthode dépasse 50 %.

- GOR : développé initialement par Garnier et collaborateurs, elle a été améliorée par Gibrat et collaborateurs [62, 65, 66, 61]. GOR I [62] utilise une approche proche de celle de Chou et Fasman, mais met en valeur les incompatibilités dans les distributions. La méthode tient compte des distribution d'angles dièdres [65, 66]. Le taux de prédiction dépasse 64 %. Plus récemment, le taux de prédiction est passé à plus de 67 % en ajoutant une information binaire dépendante juste de l'hydrophobicité du type d'acides aminés [104].

- Alignement de séquences : Du fait de l'augmentation des bases de données aussi bien protéiques que génomiques, il est possible d'utiliser des séquences similaires dans des techniques de prédiction [68, 64, 211]. Le principe est séquentiel et consiste à aligner une séquence à prédire avec d'autres séquences connues, puis à observer dans les zones consensus les compatibilités existantes pour moduler la prédiction finale. Par exemple, une procédure bayésienne qui donnait d'un taux de prédiction moyen de 67 % passe avec cette méthode à plus de 73 % de bonne prédiction [193]. Le programme PREDATOR, lui, passe d'un pourcentage de 68 % à 75 % [60]. Même un nombre limité de séquences dans l'alignement permet des gains significatifs [42].

- Il faut noter que certains algorithmes ne tiennent pas compte de la cohérence séquentielle des résultats. Ainsi une hélice peut être prédite directement aprés un feuillet sans avoir de boucles entre les deux états [210]. DSC [105] en revanche s'en préoccupe. Il utilise un ensemble de paramètres physico-chimiques et en tenant compte des insertions et des délétions présentes dans l'alignement atteint un taux de 70,1 % seul. Pour certaines protéines, il dépasse les méthodes de prédiction par réseau neuronal qui sont décrites dans le paragraphe suivant.

Ainsi, il existe un grand nombre de méthodes fort différentes basées sur des principes statistiques. L'intérêt majeur de ces méthodes, en regard des méthodes d'apprentissages type réseaux neuronaux, est la possibilité de comprendre les facteurs entrant en jeu dans la prédiction, donc les acides aminés prépondérants et leurs dépendances éventuelles.

2.2.1.2 Réseaux neuronaux et alignements de séquence

Avec l'augmentation du nombre de protéines disponibles, l'utilisation de méthodes d'apprentissage plus gourmandes en données comme les réseaux neuronaux était possible. Les premières

études n'ont utilisé comme information que les séquences *brutes* avec un réseau à une couche [148, 85, 173] ou encore avec deux réseaux imbriqués [133]. Pour une explication rapide des réseaux neuronaux, voir l'Annexe 1.

La prise en compte des alignements de séquences a permis une augmentation importante du taux de prédiction :

- PHD [166, 165], développé par Rost et Sander, est l'un des plus connus du fait d'une diffusion importante sur le Web [167]. PHD est un ensemble composé de deux réseaux possédant chacun une couche cachée. Le premier apprend la succession des acides aminés avec en plus des informations arrivant des alignements de séquences et d'échelle d'hydrophobicité. Le second réseau, lui, utilise les mêmes informations avec en plus le résultat de la prédiction et ainsi élimine les incohérences possibles de type "passage direct d'une hélice à un feuillet". Cet ensemble d'information permet de dépasser aisément les 72 % de bonne prédiction.

- NNSSP utilise un réseau avec une seule couche cachée [168]. Il a été conçu pour prendre une séquence seule ou un alignement de séquences. Le pourcentage de bonne prédiction passe alors de 71,0 % à 73,5 % en utilisant des séquences possédant des homologies. Ces résultats sont à mettre en parallèle avec l'ancienne version SSP (purement statistique) dont les taux de prédiction étaient respectivement 65,1 et 68,2 % [182].

- Des travaux montrent l'importance de la taille de la base de données impliquée dans l'apprentissage, et celui du nombre de neurones dans les couches cachées. L'optimisation de tels paramètres permet une augmentation importante du taux de prédiction de 2 à 4 % [27, 28] et, récemment, le taux de 80 % aurait été dépassé [144]. Les valeurs maximales attendues de ce type d'analyse avoisineraient les 85 % dans un futur proche [59].

2.2.1.3 Méthode consensus

Elles sont de deux types, la première consiste à utiliser à la fois en un seul algorithme différentes méthodes puis à combiner les informations. En parallèle, des méthodes statistiques et des méthodes de réseau neuronal peuvent être utilisées [187], ou encore faire une procédure plus hiérarchique où certains algorithmes sont pris au départ pour être ensuite affinés par

d'autres résultats de prédiction. Ce dernier type de méthode permet, en utilisant des méthodes statistiques et des réseaux neuronaux avec alignement de séquences, de dépasser des taux de 76 % et surtout d'obtenir des taux excellents sur les feuillets β qui sont normalement les plus difficiles à prédire [142].

Une autre possibilité est largement exploitée : l'utilisation conjointe de logiciels disponibles avec un consensus selon le type de résultats des prédictions. Ainsi SOPMA donne 69,5 % de bonne prédiction, ensuite couplée avec PHD, ce qui permet d'obtenir plus de 82 % de bonnes prédictions pour les résidus ayant été prédits par les deux méthodes dans le même état (soit près des 3/4 des résidus) [64]. La combinaison de 4 des algorithmes les plus connus (DSC [105], PHD [165], NNSSP [168] et PREDATOR [60]) permet ainsi d'atteindre 75,4 % (*http://barton.ebi.ac.uk*) [34, 33]. Actuellement, il est possible de combiner près d'une dizaine de méthodes en parallèle (*http://ibcp.fr*) [40, 72].

Ces dernières méthodes tendent aussi à démontrer que le rôle des interactions à longues distances considérées comme la cause du faible taux initial des prédictions des structures secondaires [82] a souvent été sur-estimé [56].

2.2.1.4 Bases de données

La base de données utilisée est capitale pour la validité de l'expérimentation. Ainsi, une base ne contenant que des protéines avec des taux d'hélices α importants fera que la méthode de prédiction sera particulièrement appropriée pour des protéines ayant de forts taux en hélices α, mais le résultat pour une protéine β sera peu probant. Cuff et collaborateurs récapitulent ainsi les problèmes liés aux premières prédictions. Faits sur des bases de données peu importantes, les taux de prédiction annonçés étaient compris entre 63 et 77 %. Des tests effectués sur de nouvelles protéines montrent en réalité des taux compris entre 50 % et 56 % [33].

Aussi, ne faut-il pas utiliser des bases complètes, mais des bases "nettoyées". Ces bases doivent satisfaire à deux critères : (i) être suffisament peuplées pour représenter la spécificité du plus grand nombre de repliements possibles, sans déséquilibre particulier, et (ii) être dépourvues de biais en terme de séquence, de façon à optimiser la détermination de la relation séquence-structure.

Une approche intéressante a été développée en Finlande [11], mais restée sans suite (sans doute du fait de sa non-disponibilité sur le web). La procédure est de type hiérarchique. Elle

utilise la séquence pour avoir un taux d'identités faible, mais aussi la structure en définissant des taux de structures répétitives par DSSP [100], pour ne pas avoir des protéines trop reliées entre elles. Les deux bases de données "nettoyées" sont PDBselect et HSSP. La première est un ensemble de protéines possédant un taux d'identité de séquences inférieur ou égal à un seuil choisi (*http://swift.embl-heidelberg.de/pdbsel/*) [84, 83]. Aucun lien avec la structure n'est testé. La méthode est basée sur l'exclusion des séquences proches jusqu'à obtention d'un taux maximal. La seconde prend en compte la structure et se base principalement sur des alignements de séquences [174, 43]. Ainsi, des séquences proches sont exclues et donc des repliements proches normalement aussi.

2.2.2 Modèles protéiques

Les structures secondaires donnent une idée de la topologie de la protéine. Il convient néanmoins d'utiliser d'autres techniques pour aboutir à la structure réelle de la protéine. Deux types de techniques existent: celles qui se basent sur l'utilisation de fragments ou de séquences protéiques connus, pour la modélisation par homologie et pour l'enfilage (ou *threading*) et celles qui se basent sur une représentation simplifiée des acides aminés pour la modélisation *ab initio*.

2.2.2.1 Modélisation par homologie

La modélisation par homologie se base sur un concept simple. Des séquences proches ont des repliements proches, comme on le constate pour des protéines ayant des taux de similitude de séquence supérieures à 30 %.

Plusieurs équipes ont montré qu'à partir de structures connues, il est possible de construire pour une séquence, ayant des taux de similitudes de séquences significatifs, de manière supervisée, une nouvelle structure avec de bons résultats [99, 30].

Au vu de la complexité et du nombre croissant de protéines, l'utilisation des procédures automatisées est devenue incontournable [120].

Différents programmes existent maintenant comme COMPOSER [10, 9] ou MODELLER [169], prenant en entrée une séquence et pouvant donner en sortie un certain nombre de modèles. Le principe de ces méthodes se rapproche fortement de la méthode développée par le groupe

de David Eisenberg, qui peut être subdivisée en 4 parties [15, 121] :

1- Classer une base de données en fragments protéiques associés à certaines caractéristiques telles que la structure secondaire, l'exposition au solvant, la position par rapport aux positions N et C-terminales de la protéine,...

2- Un profil est effectué sur des alignements de séquences [71]. Ce profil est alors comparé à la base de données et l'on cherche les compatibilités locales de manière extensive.

3- Les parties associées aux hélices et aux feuillets sont trouvées en premier. La construction des zones variables est plus complexe. Une recherche de compatibilité entre la disposition spatiale et les contraintes des profils est effectuée.

4- Les chaînes latérales sont ajoutées et l'ensemble est minimisé. Des ensembles de rotamères sont ici utilisés. Ils représentent des prototypes moyens des conformations des chaînes latérales. Pour choisir lequel est associé à un résidu, l'interaction avec les voisins et les contraintes géométriques sont déterminées.

2.2.2.2 Technique d'enfilage (ou *threading*)

Les techniques d'enfilage (ou *threading*) sont utilisées principalement dans le cas de très faibles homologies et se basent sur une étude énergétique [54, 95, 17, 126]. Construisant la chaîne polypeptidique résidu par résidu, on cherche à l'allonger au mieux en calculant les incompatibilités existantes entre la séquence utilisée et des bases de données de structures cristallographiques. Diverses fonctions d'énergies sont alors requises. Les résultats étaient initialement fort dépendants de la taille et surtout de la composition en structures secondaires répétitives [96]. Ainsi THREADER 2 obtient des résultats particulièrement intéressants dans la reconnaissance des repliements dans la série CASP (pour *Critical Assessment of Techniques for Protein Structure Prediction*, ensemble de congrès où les méthodes sont testées sur des nouvelles séquences dont la structure réelle n'est révélée qu'à la fin) [97]. Outre une difficulté classique de paramétrisation des champs de force, la difficulté de la méthode vient de l'incertitude des régions d'insertions. Différentes améliorations existent pour moduler ce problème comme les techniques d'enfilage à "champ de force auto-consistant" [55] où le champs de force est ajusté à la séquence examinée et non pas à celle de la séquence cible.

Une autre difficulté est l'obtention de nombreux modèles et donc la recherche du "meilleur" [152]. L'utilisation de séquences homologues aide particulièrement quand il n'y a pas d'insertions dans l'alignement [153], du réseau neuronal PHD en parallèle [7] ou des méthodes statistiques comme les Chaînes de Markov Cachées [8]. Le temps de calcul est souvent loin d'être négligeable [205].

Les méthodes qui semblent les plus prometteuses sont des méthodes hiérarchiques, tel LINUS [183] qui part d'un modèle et le raffine en recherchant avec une minimisation rapide les zones qui sont apparemment associées avec un type de repliements. Il prend en compte différentes méthodes dans un ordre précis [93] ou encore utilise des données phylogénétiques [143].

2.2.2.3 Modélisation *ab initio*

Enfin, une dernière technique est applicable, la modélisation *ab initio*. Elle consiste en une représentation schématique des protéines. Le plus souvent un résidu n'est représenté que par un atome, souvent le C_α, et, avec différents jeux de contraintes le système évolue pour explorer au maximum l'espace conformationnel [41]. Les pseudo-atomes sont déplacés suivant des mouvements tenant compte de processus aléatoires ce qui permet de dépasser certains minima locaux. Il peut y avoir des contraintes énergétiques liées à des contraintes de distances [88], des contraintes liées à la séquence à un niveau local par des méthodes statistiques [179, 180, 178], avec des prédictions de structures secondaires [137] ou encore en recherchant dans des bases de données [38]. Des techniques combinant plusieurs approches permettent d'avoir une idée plus précise du type de topologie de la protéine [171].

Toutefois, les résultats sont fortement liés à la taille de la protéine ainsi qu'au type de protéines [39, 185, 4, 141].

2.2.2.4 Quelques exemples de prédiction jouant un rôle important dans le repliement protéique

La nature des acides aminés fait que certains résidus aiment plus ou moins le solvant, et qu'ils préfèrent interagir avec certains autres acides aminés plutôt que d'autres. Ces propriétés physico-chimiques induisent leur capacité à créer des contacts entre eux, ce qui induit au final la structure 3D de la protéine Ainsi, trois catégories de prédictions importantes dans la modélisation moléculaire sont :

chaînes latérales. Elles se basent sur une recherche locale du minimum énergétique [197, 14] par une utilisation successive des différents rotamères possibles. Les rotamères sont des prototypes moyens de chaînes latérales [196, 45, 195].

accessibilité. Le calcul de l'accessibilité se fait le plus souvent en faisant rouler une bille virtuelle sur la protéine [118, 91, 155]. La prédiction de l'accessibilité permet de rendre compte du rôle prépondérant de l'effet hydrophobe/hydrophyle [129, 12, 25]. Le score maximal de prédiction est de 77,6 % pour un taux relatif de 9 % d'accessibilité [128]. A plus de 90%, c'est l'effet hydrophobe/hydrophyle qui joue le rôle primordial [130]. L'accessibilité est d'un point de vue biologique capital, car l'ensemble des fonctions biochimiques des protéines sont assurées à la surface des protéines. C'est le site de liaison des autres composés protéiques [32] ou nucléiques [92, 134].

contacts. La prédiction des contacts à partir de la séquence primaire est un des défis les plus difficiles actuellement. Les meilleurs algorithmes ne dépassent pas 15 % de bonnes prédictions [128, 48]. Certains algorithmes s'intéressent au point précis de contacts, d'autres à des zones plus étendues [135, 124]. Un des problèmes majeurs est lié à la recherche à "longue distance" (plus de 40 résidus), en dehors des ponts disulfures; les prédictions sont à peine plus fiables qu'une recherche aléatoire [67].

Aussi, d'autres types d'informations plus simples sont exploitées comme prédire le nombre total de contacts dans une protéine [49], ou encore les Cystéines impliquées dans une liaison disulfure. Dans ce dernier cas, les taux de prédiction dépassent les 70 % [50, 57].

2.3 Les alphabets structuraux

2.3.1 But

Les structures secondaires permettent une description structurale précise des deux structures périodiques que sont les hélices α et les feuillets β (cf. paragraphe 2.1.2.1). Toutefois, la classification à l'aide des structures secondaires laisse un peu plus de 50% des structures protéiques non attribuées (les boucles). La description structurale est finalement assez limitée. La construction de bibliothèques de boucles définies par rapport aux structures répétitives apporte

une aide dans l'approximation de la structure et permet d'établir des classifications utilisables dans des méthodes de prédiction [44, 204]. Les boucles de taille supérieure à 8 ne sont cependant jamais prises en compte dans ces approches. La reconstruction à partir de la seule information des structures secondaires ne permet pas une reconstruction du repliement 3D.

Aussi, de nombreuses études ont été menées pour tenter de catégoriser et de classifier des fragments protéiques de petites tailles (inférieures à 20 acides aminés) représentatifs des protéines présentes dans les bases de données de structures cristallographiées, la PDB. Les différentes méthodes présentées ici cherchent à examiner des successions de quelques acides aminés, les structures statistiquement répétées les plus courantes, pour décrire au mieux l'ensemble des repliements des structures protéiques.

Le travail précurseur de Jones et Thirup doit-être noté ici. Ils ont reconstruit une protéine de liaison au rétinol avec des fragments structuraux tirés de 3 protéines ayant peu d'homologie de séquence. Ce travail se basait sur l'étude d'une carte de densité électronique [99]. Depuis, deux approches assez distinctes se trouvent dans la littérature. La première cherche un nombre important de prototypes, une centaine pour représenter au mieux les protéines [198, 146, 200, 175]. La seconde recherche un nombre plus limité de prototypes, normalement inférieur à une vingtaine, afin de permettre leur utilisation dans des méthodes de prédiction plus fiables [162, 163, 53, 19, 23, 24]. Je parlerai de blocs structuraux ou parfois de blocs protéiques dans cette partie, les notions étant équivalentes et correspondant toujours aux petits prototypes, nommés *Buildings blocs (BB)* par Unger et collaborateurs [198], *Short Structural Motifs (SSM)* par Unger et Sussman [200], *Substructure* par Prestrelski et collaborateurs [146], *Local Structural Motifs (LSM)* par Schuchhardt et collaborateurs [175], *Recurrent Local Structural Motifs (RLSM)* par Rooman et collaborateurs [162, 163], *Structural Buildings blocs (SBB)* par Fetrow et collaborateurs [53], *Local Structure (LS)* par Bystroff et Baker [19], et les *Short Structural Building blocs (SSB ou SSBB)* de Camproux et collaborateurs [23, 24].

Dans les pages qui vont suivre, je vais présenter les différentes études qui ont été menées à l'heure actuelle en développant la méthodologie et les résultats propres à chaque équipe. Avec dans un premier temps, les études sur des nombres de blocs importants (plus de 100), ensuite ceux sur les nombres faibles de blocs (entre 4 et 13), puis enfin le travail du groupe de Baker, qui est plus complexe.

2.3.2 Un nombre important de blocs protéiques pour une description fine de la structure protéique

2.3.2.1 Par une méthode d'aggrégation en deux étapes (Unger *et al.*, 1989, 1993)

L'objectif de ce travail est d'obtenir un grand nombre de prototypes pour pouvoir ensuite reconstruire l'ensemble des structures protéiques avec ces blocs (1989,[198] et 1993,[200]). Pour leur base de données, Unger et collaborateurs ont conservé sur les 354 chaînes présentes (PDB de janvier 1987), 82 ayant un facteur R dit "correct" et une résolution de moins de 2.8 Å, soit 12 973 résidus.

Leur méthode consiste à calculer l'écart quadratique moyen ou *RMSd* (root mean square deviation) entre deux structures s et t, en ne conservant que les carbones α du squelette dans cette comparaison:

$$RMSd(s,t) = \sqrt{\frac{\sum_{i=1}^{n}(r_i^s - r_i^t)^2}{n-2}}$$

Un exemple intéressant des problèmes liés au calcul du *RMSd* est donné. Le *RMSd* minimal entre deux structures n'est pas obligatoirement une information adéquate. Ainsi la figure 2.3.2.1 montre 3 fragments protéiques. Le *RMSd* entre (a) et (b) est plus grand que celui entre (b) et (c), alors que le type de repliement de (a) est plus proche de celui de (b).

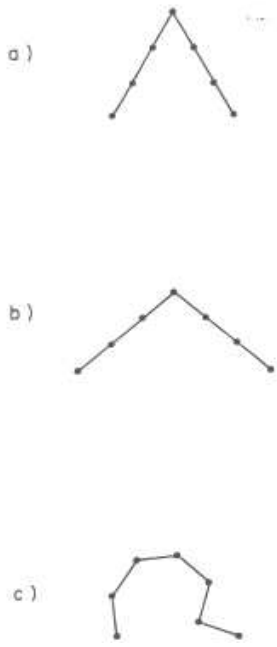


FIG. 2.9 – *3 fragments protéiques d'une longueur de 7 C_α (Figure 1. p.357 [198]).*

Après un calcul préliminaire effectué sur 4 protéines (426 résidus au total), Unger et collaborateurs ont décidé de prendre comme prototypes des hexamères (i.e., des fragments de longueur 6). Ils considérent cette longueur comme la plus petite correcte pour bien différencier les fragments protéiques entre eux. L'ensemble des *RMSd* entre les hexamères des 4 protéines en question (soit 82 215 couples) est calculé et ensuite réuni par une méthode dite "d'annexion".

Cette procédure se passe en trois étapes:

1- Un hexamère est tiré au hasard dans la base de données. Tous les hexamères étant à moins de 1 Å de *RMSd* de cet hexamère lui sont attribués. Ensuite tous les hexamères du groupe sont utilisés pour leur associer les hexamères étant à moins de 1 Å. La procédure s'arrête quand plus aucun hexamère ne peut être attribué au groupe. Un hexamère n'appartenant à aucun groupe est alors tiré au sort et le processus recommence jusqu'à l'attribution de tous les hexamères à un groupe.

2- L'inconvénient de la première étape est de créer des groupes importants dont certains ont plus d'1 Å de RMSd. Aussi, Unger et collaborateurs redivisent les groupes et font plusieurs tentatives pour n'obtenir que des sous-groupes dont tous les hexamères sont à moins de 1 Å de *RMSd* les uns des autres.

3- Le centre de chaque groupe est conservé comme bloc structural type du groupe (bloc le plus proche du barycentre calculé).

La méthode d'annexion est moins conventionnelle que la méthode des nuées dynamiques. La plus grande crainte des auteurs est de n'obtenir que fort peu de groupes dans la première étape. Elle donne 55 groupes, ce nombre étant directement lié au choix de leur valeur limite de 1 Å , qui est assez faible pour des fragments de cette taille. La seconde étape est plus criticable. En effet, rien ne laisse supposer qu'un groupe X doit être subdivisé en sous-groupe X_1 et X_2. Il est logique de penser que des éléments de X après division puissent être alors plus proches d'un autre groupe Y plutôt que X_1 ou X_2. Après la seconde étape, 103 blocs structuraux sont obtenus. Sur la base de données complète :

a- 76% des fragments sont proches d'un des blocs structuraux avec un *RMSd* de moins d'1 Å.

b- 92% avec moins de 1,25 Å.

c- 65% ne sont proches que d'un bloc (à moins d'1 Å).

d- 5% sont proches de plus de 2 blocs (à moins d'1 Å).

e- En ne conservant que des fragments ayant moins de 1 Å de différence avec la réalité, 99% de leur base de données est couverte.

Utilisant 4 autres protéines, la méthode donne à 144 blocs protéiques finaux. En combinant les 8 protéines, elle atteind à 170 blocs structuraux. Les plus fréquents sont retrouvés dans les deux expériences, mais le reste est trés fluctuant. Les structures connues répétitives et leurs liens caractéristiques sont retrouvés.

Une méthode de reconstruction sur des protéines de longueur 60, ou alors sur les 60 premiers acides aminés (extrémité N-terminale), est effectuée en ajoutant à chaque fois "au mieux" un nouveau carbone correspondant au nouveau bloc protéique. La procédure consiste à superposer en une position X les 5 premiers atomes du bloc structural x avec les 5 derniers de la chaîne reconstruite. La reconstruction est correcte dans plus de 64% des cas. Seuls des changements rapides dans le repliement posent un réel problème.

En conclusion, la méthode développée par Unger et collaborateurs est intéressante car elle est basée sur une méthodologie pertinente pour regrouper des fragments protéiques, et surtout, les blocs ont servi lors d'une méthode de reconstruction 3D avec de bons résultats. Toutefois, deux points posent problémes :

(a) Le choix de 4 protéines seulement au départ est trop faible et ne peut rendre compte de la diversité des repliements protéiques, comme le nouvel échantillonnage le montre fort bien. Un choix de 25 protéines soit 1/3 de la base de données aurait été plus judicieux. Pour éviter un temps de calcul élevé, un traitement rapide pour éliminer du regroupement des fragments vraiment trop proches, comme des coeurs d'hélices ou de feuillets, aurait été possible.

(b) La méthode de réallocation est fortement discutable, la taille des groupes devenant rapidement très faible. Par exemple, le seizième groupe le plus important de la base de données représente 1.0%. Sur la base de données de départ de 4 protéines, il ne représente donc que 4 fragments.

L'équipe d'Unger utilisera ces blocs, comme exemple, pour une autre méthode de reconstruction, basée sur les angles dièdres [199]. Ils montrent que la plupart des hexamères trouvés sont créables par une procédure aléatoire dans les zones de Ramachandran classiques.

Enfin en 1993, Unger et Sussman réutilisent les blocs obtenus quatre ans auparavant [200], mais n'en conservent que 81, ceux-ci étant observés plus de 35 fois dans leur base de données, soit une fréquence minimale de 0,3 %. Détail troublant, les chiffres donnés sont les même que ceux obtenus pour les 103 blocs. Les hexamères obtenus se retrouvent dans les zones préférentielles du diagramme de Ramachadran, mais beaucoup sont significatifs car une procédure purement aléatoire crée beaucoup de fragments peu ou non observés dans la base de données. Ainsi, les auteurs espèrent déceler un début d'architecture des protéines avec leurs hexamères. Pour explorer un cas plus localisé, tous les blocs définis en 'brin étendu' ('E' ou feuillet β dans DSSP [100]) sont extraits de la base de données. Ces différents blocs ont une répartition différente en acides aminés. Toutefois, les chiffres donnés sont en pourcentage et on ne peut juger de la valeur statistique des variations observées.

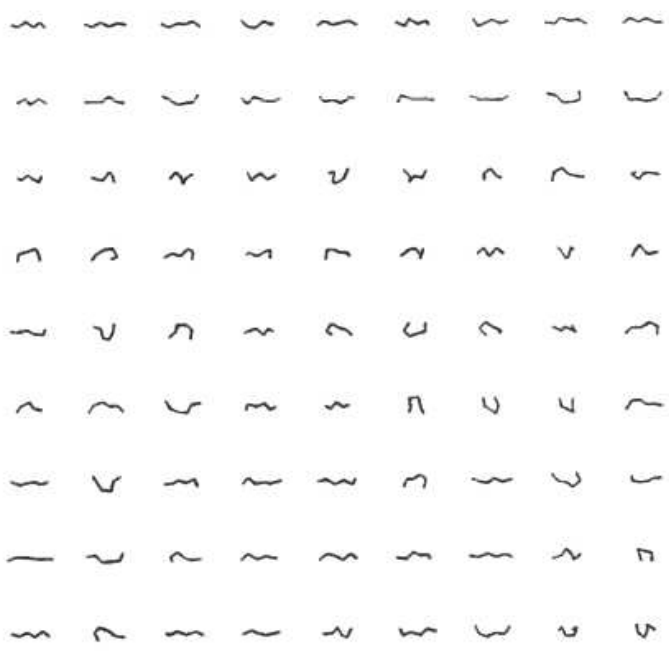


FIG. 2.10 – *Représentation 3D des 82 blocs structuraux (BB) (Figure 1. p.463 [200]).*

2.3.2.2 Pour une librairie de sous-structures (Prestrelski *et al.*, 1992)

Prestrelski et collaborateurs ont créé une librairie de blocs, sans *a priori* sur le type de structures secondaires. Ils désirent les utiliser pour trouver des homologies structurales, d'une

manière similaire à celle de Jones et Thirup [98], mais en concevant une librairie fixe de prototypes et non spécifiquement conçue pour une protéine unique (1992, [146]).

La base de données utilisée est composée de 14 protéines ne possédant pas d'homologie structurale, résolues à moins de 2,5 Å et représentant 2 437 résidus.

La méthode consiste en la génération d'un petit nombres de blocs différents structuralement. Le critère de similarité choisi est moins classique que dans le précédent travail. Il s'agit d'une fonction liant un critère de distance dite distance linéaire et un critère angulaire, l'angle α qui décrit 4 C_α successifs du squelette polypeptidique. La figure 2.11 montre ces deux critères.

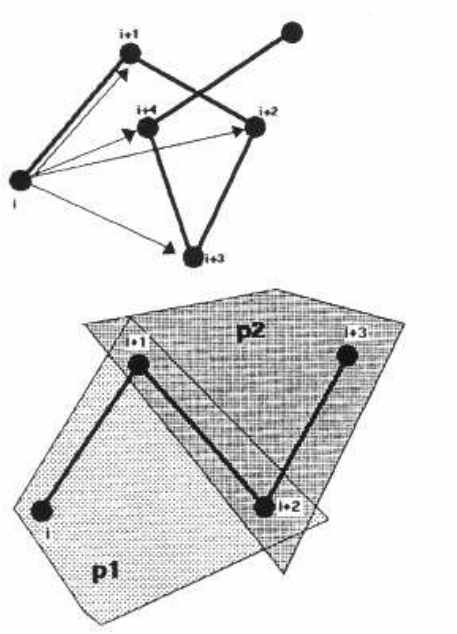


FIG. 2.11 – *Représentation schématique de la distance linéaire et de l'angle α [146].*

La distance linéaire (DL) est la somme des distances C_α-C_α [122]:

$$LD_i = \sum_{j=1}^{4} (\mathbf{d}_{i,i+j})$$

avec $\mathbf{d}_{i,i+j}$ la distance entre le $C\alpha_i$ et le $C\alpha_{i+j}$. Ce critère permet de différencier les structures répétitives. Elle a été utilisée pour observer des insertions-délétions dans des structures cristallographiques. Toutefois, elle ne permet pas de différencier une hélice gauche d'une hélice droite par exemple [122]. Pour contrecarrer ce type de problème, ils ont donc adjoint l'angle α. Pour différencier deux fragments protéiques s et t, une fonction de coût $C_{s,t}$ dépendant de deux paramètres C_1 et C_2 a été définie:

$$C_{s,t} = C_1 \times (|DL_s - DL_t|) + C_2 \times (\tan|\alpha_s - \alpha_t|) + C_2 \times (\tan|\alpha_{s+1} - \alpha_{t+1}|)$$

La valeur maximale de la différence des angles α est limitée à 85^o. Pour travailler sur des fragments de taille supérieure à 5, ils ont utilisé une fenêtre glissante définie par $Max = (\mathbf{L}_f$ - 4 $)\times$ $maxdif$ $\times 0,75$, avec $\mathbf{L}_f$ la taille du fragment considéré, $maxdif$ la valeur maximale de la fonction de coût pour les fragments considérés. La valeur 0,75 sert à comparer des fragments de taille 8 utilisés dans l'étude et ceux de taille 5.

Quand les valeurs des distances linéaires sont proches, l'angle α est très différent. L'utilisation conjugée des deux critères semble donc discriminante.

Avec cette approche, la première étape à réaliser est la calibration des deux coefficients C_1 et C_2. Pour les ajuster, ils ont testé un ensemble de valeurs de C_1 allant entre 0,1 et 1,0 et de C_2 entre 1,0 et 2,2. Pour déterminer la qualité discriminante des coefficients, ils ont testé une dizaine de prototypes d'hélices, de feuillets et de structures périodiques en calculant leurs distances linéaires, leurs angles α et les valeurs de Max associées aux coefficients testés. Ensuite, ils ont choisi comme notion d'équivalence le fait que deux structures ont un $RMSd$ inférieur à 1 Å. Les résultats ont particulièrement pris en compte la distinction entre les hélices α et le reste des exemples. C_1=0,9 et C_2=2,0 ont ainsi été choisis et appliqués comme critère de coût à l'ensemble des fragments issus de la base de données.

La méthode utilisée de groupements n'est pas décrite en détail, mais aux vues des résultats donnés qui récapitulent les 30 blocs les plus fréquents, une simple recherche des coûts minimaux aurait pu suffir. Ils obtiennent au final 113 blocs distincts. Pour donner un ordre de grandeur, le deuxième bloc le plus fréquent ne représente que 17 fragments.

En conclusion, la méthodologie utilisée est peu conventionnelle. Les seuls points ambigus concernent la recherche des coefficients C_1 et C_2 et la signification exacte du coût pour les fragments de longueur 8 usités. Si la formule donnée est bien celle utilisée, le seul critère important est $maxdif$ qui représentent en fait $argmax = \{C_{s,t}\}$ pour chacuns des fragments, s et t. Ceci convient parfaitement pour "écarter" les fragments. On ne travaille que sur les différences maximales. Toutefois, il est possible que cette fonction génère un nombre de blocs distincts trop important. Avoir plus de 72% des blocs vus moins de 10 fois paraît un peu

élevé. Unger et collaborateurs ont des chiffres fort différents. Un autre facteur a pu jouer: le choix des protéines de la base de données qui sont sans homologie structurale. Au final, seules les hélices α réapparaissent fortement; les feuillets β sont peu présents. Cela peut être dû à leur caractéristique structurale plus lâche. Par ailleurs, les auteurs désirent faire plutôt une classification par classe de protéines. L'adjonction de 6 nouvelles protéines fait passer leur taux de recouvrement par les 30 prototypes les plus courants de 67% à 77%.

Cette méthode a été mise au point pour aider lors de travaux de spectroscopie à infra-rouge. La librairie générée leur a permis de constater des corrélations impossibles avec une autre méthode et donc de proposer une architecture de sérine-protéase compatible avec les structures connues [147].

2.3.2.3 Par une Carte-auto organisée de Kohonen (Schuchhardt *et al.*, 1996)

L'objectif des travaux de Schuchhardt et collaborateurs est d'obtenir un grand nombre de prototypes pour approximer la structure tridimensionnelle d'une protéine dans un but de classification et d'analyse sans tenter de reconstruire cette structure (1996, [175]).

La base de données est composée de 136 protéines ayant moins de 30% de similitude de séquence [84, 83], ce qui représente un total de 24 239 résidus.

La méthode consiste à caractériser les protéines en utilisant des fragments d'une longueur de 9 résidus décrits par les angles ϕ et ψ. Les protéines sont découpées en fragments de 9 résidus donnant chacun un vecteur d'observation de longueur 16 (le premier ϕ et le dernier ψ n'étant pas pris en compte). L'ensemble comporte donc 23 151 exemples. Schuchhardt et collaborateurs ont utilisé une méthode non supervisée d'apprentissage, les cartes de Kohonen (cf. Annexe 1 pour plus de détail). Cette méthode consiste à créer $N \times M$ neurones (représentés par une matrice), chaque neurone étant une observation moyenne, ici un vecteur d'angles dièdres, à rechercher pour chaque observation de la base de données le neurone le plus proche de cette observation et lui faire apprendre cette observation. Le principe est décrit en détail dans l'annexe I. Son intérêt principal est un aspect visuel particulièrement utile pour l'analyse. La méthode des cartes de Kohonen ne diffère des nuées dynamiques ou des k-moyennes (ou *k-means*) [79] que par l'existence d'un processus de diffusion qui permet de réunir dans un environnement proche des neurones ayant des points communs.

Pour évaluer les différences entre les observations, une mesure de dissemblance, le *RMSda*

ou *root mean square deviation on angular values* est utilisée. Cette distance sur les angles a déjà servi dans une méthode de recherche d'homologie structurale [102]. Il s'agit d'une distance euclidienne entre deux vecteurs d'observations s et t:

$$RMSda(s,t) = \sqrt{\frac{1}{16}\sum_{1}^{8}(\phi_s^i - \phi_t^i)^2 + (\phi_s^{i+1} - \phi_t^{i+1})^2}$$

La figure 2.12 montre la carte de Kohonen obtenue pour une taille de 10×10, soit 100 neurones. Cette carte est "plane", ce qui veut dire qu'elle possède des bords, le neurone (9/9) n'est voisin que des neurones (8/8), (8/9) et (9/8), alors qu'au milieu de la carte, un neurone est entouré de 8 voisins.

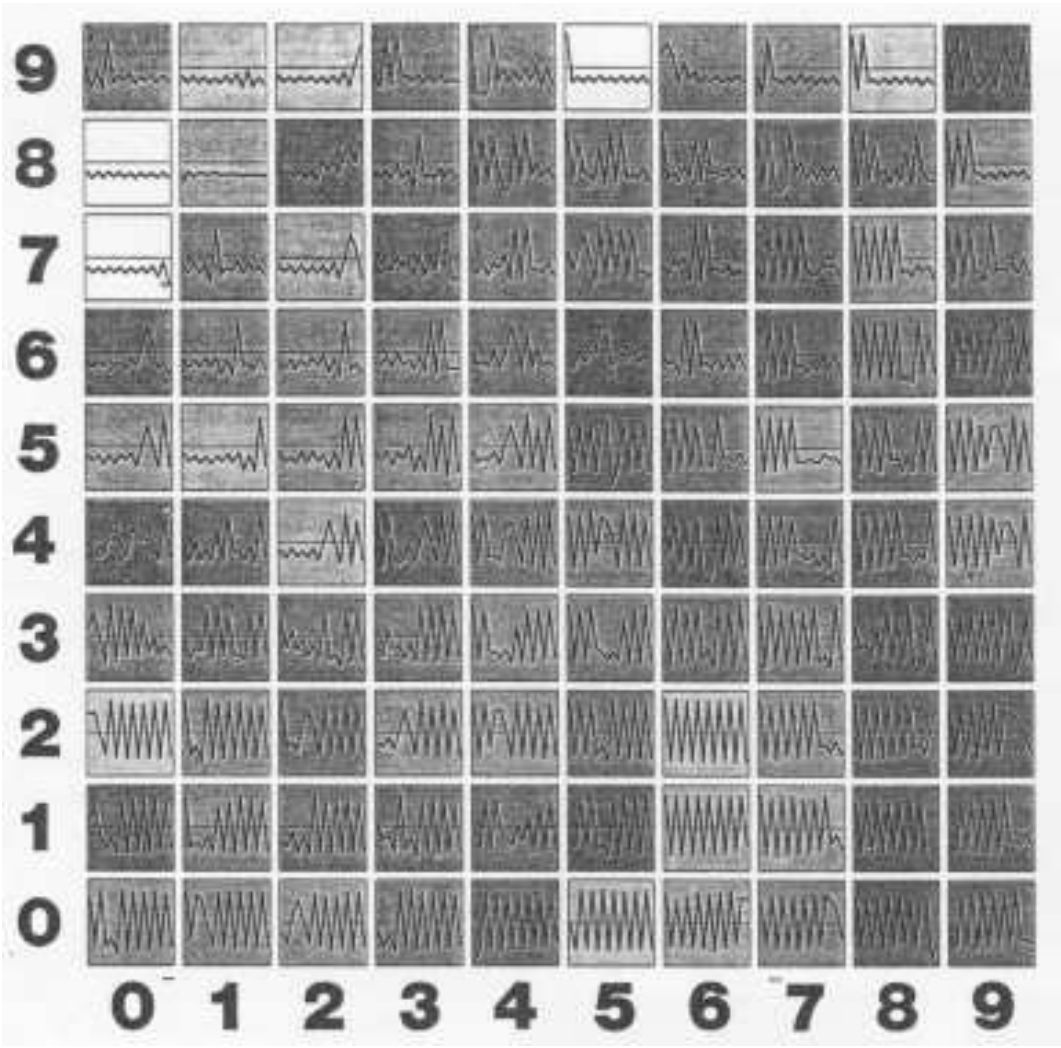


FIG. 2.12 – *Carte auto-organisée de Kohonen obtenue pour 100 neurones (Figure 4a. p.836 [175]). Chaque neurone est décrit par la suite des 16 angles dièdres moyens.*

Les feuillets et les hélices se retrouvent dans des neurones distincts. Les hélices α sont forte-

ments localisées en haut et à gauche de la carte, les feuillets β plus sur la droite. La validation structurale est effectuée exclusivement sur le critère du *RMSda* et les résultats semblent satisfaisants. Des motifs dits de "cassure" classique avec une forte proportion de Glycine sont trouvés. Quelques exemples de suivi de trajectoires de protéines sur la carte sont montrés.

La méthode est particulièrement appropriée à la recherche et à l'analyse de données biologiques, comme le montre le travail de Hanke et Reich sur la reconnaissance de motifs propres à des familles protéiques [77, 78]. L'utilisation des angles dièdres avec le *RMSda* est beaucoup plus souple que le classique *RMSd*. Du fait de sa rapidité, un travail plus important sur les différents paramètres de l'apprentissage est possible. Les résultats sont convaincants. Toutefois, ayant refait leurs expériences, deux points me paraissent importants :

1- L'utilisation d'un réseau plan est peu approprié. Un réseau fermé, où les neurones ont tous 8 voisins (i.e, le neurone (9/9) est voisin des neurones (8/8), (8/9), (9/8) et (0/0), (0/8), (0/9), (9/0) et (8/0)), permet une meilleure diffusion sans effet de bord. Et, en fin d'apprentissage, les hélices et les feuillets sont bien mieux séparés. Les motifs caractéristiques sont aussi plus accentués.

2- Une longueur de neuf résidus est un peu élevée. La recalcul des angles dièdres réels à partir d'une longueur est peu évidente. Cette méthode est donc peu utilisable en particulier pour reconstruire une protéine. De plus, la précision obtenue est parfois trés médiocre pour certains neurones, d'ailleurs aucune notion de variabilité n'est donnée dans le texte. Certains angles ont une variabilité proche de 180^o. Avec une carte plus petite, on conserve un approximation comparable, et une longueur plus faible, elle est plus intéressante, en fait.

2.3.3 Peu de blocs protéiques

2.3.3.1 Utilisation d'un regroupement hiérarchique (Rooman *el al.*, 1990)

Les travaux de Rooman et collaborateurs ont consisté en l'obtention d'un petit nombre de blocs structuraux, représentant pertinement la base de données. Ces blocs sont de taille fixe, mais plusieurs longueurs ont été étudiées (1990, [162]). Ils ont été ensuite utilisés dans une méthode de prédiction de la structure protéique à partir de la séquence; les auteurs ont surtout

	N.obs.	freq. rel.	*RMSd* moyen (Å)
η_4	4858	38,1	0,33
ϵ_4	1795	14,1	0,24
ζ_4	3151	24,7	0,39
λ_4	2949	23,1	0,60
η_5	3917	31,0	0,48
ϵ_5	2639	20,9	0,44
ζ_5	3377	26,7	0,82
λ_5	2700	21,4	1,07
η_6	3645	29,1	0,76
ϵ_6	4845	38,7	1,00
ζ_6	1246	10,0	1,08
λ_6	2779	22,2	1,32
η_7	3997	32,2	1,14
ϵ_7	3469	27,9	1,06
ζ_7	2652	21,5	1,67
λ_7	2284	18,4	1,56

TAB. 2.7 – *Nombre d'observations (N.obs.) pour chaque bloc obtenu, avec sa fréquence relative (freq. rel.) et le* RMSd *moyen (Figure3. p.332-333 [162])*

mis en parallèle ces blocs et les structures secondaires (1990, [163]).

La base de données utilisée se compose de 75 protéines ayant une bonne résolution et possédant peu de similitude de séquence. Cela représente à 12 978 résidus.

La méthode d'apprentissage se base sur une classification hiérarchique qui utilise comme critère de distance entre fragments protéiques le *RMSd* (cf section 2.3.2.1). Dans un premier temps, les fragments de taille L sont tous comparés deux à deux grâce au *RMSd* sur les C_α du squelette polypeptidique. Ensuite, une classification hiérarchique est effectuée.

Des longueurs L allant de 4 à 7 C_α ont été testées, et, pour chaque longueur, 4 groupes distincts ont été déterminés. La table 2.7 récapitule le nombre d'observation et le *RMSd* moyen de chaque groupe. J'ai ajouté la fréquence relative pour avoir une idée de la contribution de chaque bloc.

Sur les 4 classes observées pour chaque longueur, la famille λ représente toujours un ensemble exclusivement boucles, la famille η étant composée d'hélices α, la famille ζ de boucles et de feuillets β et la famille ϵ de feuillets β. Cette classification hiérarchique est utilisée ensuite pour discriminer différentes familles protéiques. La figure 2.13 donne les valeurs des angles ϕ et ψ obtenues. Je les ai caculées à partir de la figure 3 p.333 [162] pour pouvoir comparer mon alphabet avec le leur.

Dans un second temps [163], le lien existant entre ces structures et les séquences a été recherché. Pour cela, un travail important de formalisme de la significativité d'une séquence a

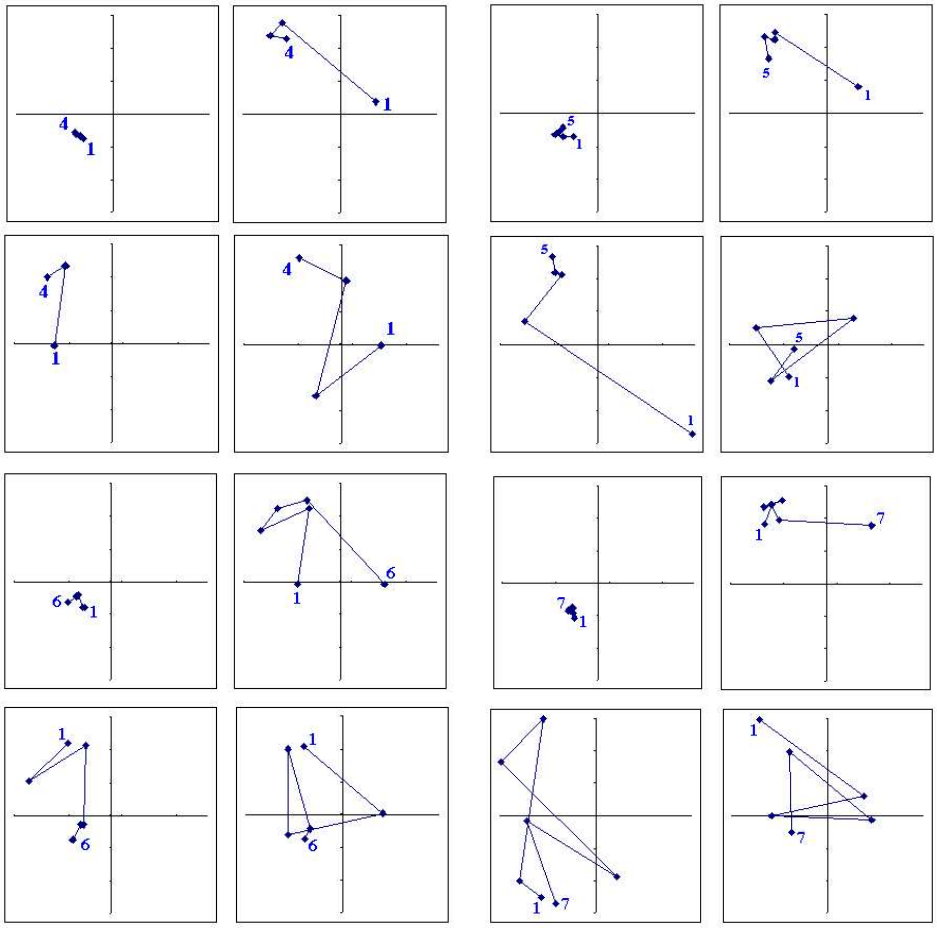


FIG. 2.13 – *Représentation des angles des 4 blocs de (a) longueur 4, (b) longueur 5, (c) longueur 6 et (d) longueur 7 (recalculés à partir de la figure3. p.331-332 [162]).*

été effectué. Chaque succession de 7 résidus a été traitée sous la forme $x-my-nz$, avec x, y et z des acides aminés déterminés compris dans une séquence de longueur 7, avec $n+m$=4 (7-3), n et m variant donc, entre 1 et 3, et représentant n'importe quel acide aminé. Les occurrences conservées sont représentatives à la fois de la séquence (un nombre d'occurrence supérieur à trois) et la structure (associée majoritairement à un type de structure donnée). Ce type de représentation avec des acides aminés non déterminés vient de la taille limitée de la base de données [164] et d'un précédent travail sur certaines formes de coudes [158].

Avec cette approche, les auteurs trouvent à partir de la séquence, un taux de prédiction compris entre 41% et 47%. Ce taux est à mettre en comparaison avec les taux des structures secondaires de l'époque qui avoisinent 60% pour 3 états [62, 65], alors que les blocs en proposent 4. En parallèle de cette observation, ils trouvent un plus grand déterminisme dans les séquences observées avec leurs 4 blocs pour une longueur L=6, mais en contre-partie un bruit de fond, lui aussi accentué. Aucune amélioration de la prédiction n'est donc observée.

Ainsi, les blocs obtenus sont fortement liés aux structures secondaires répétitives. Le choix du nombre final de blocs peut porter à critiques. Il ne permet pas, comme pour les structures secondaires, de reconstruire directement une structure protéique à partir de cette information. La taille de la base de données est pour beaucoup dans le taux final de prédiction.

Par la suite, seule la méthode de prédiction, mais appliquée aux structures secondaires, sera réutilisée [159] et mise en oeuvre dans une recherche basée plus spécifiquement pour une recherche à partir de coordonnées extraites de 7 régions dans la carte de Ramachandran [159, 160] et avec utilisation de protéines homologues [161].

2.3.3.2 Utilisation d'un réseau de neurones (Fetrow *et al.*, 1993, 1997)

L'objectif de ce travail est de reclassifier les structures protéiques en "super" structures secondaires avec des blocs obtenus à l'aide d'une méthode de compression de l'information, puis de regroupements des données ainsi traitées en quelques prototypes (1993, [208] et 1997, [53]).

Plusieurs bases de données ont été utilisées. La plus ancienne [208, 209] est composée de 74 chaînes ayant un taux d'identité de séquence inférieur à 50%, une résolution de moins de 2,5 Å, et, un facteur R inférieur à 0,3. La base est composée de 13 114 acides aminés. La plus importante, et plus récente [53], possède moins de 25 % d'identité de séquence pour 116 chaînes

protéiques représentant 23 355 résidus.

La méthode repose sur un réseau de neurones dit réseau auto-associatif (autoANN). Le principe, exposé dans la figure 2.14, est de prendre un vecteur d'information en couche d'entrée, de le comprimer dans la couche cachée comme pour tout réseau neuronal artificiel classique, mais au lieu d'avoir en couche de sortie une information nouvelle tirée des observations mises en couche d'entrée, on recherche la même information que les observations mises en entrée. En résumé, le réseau doit restituer ce qu'il voit.

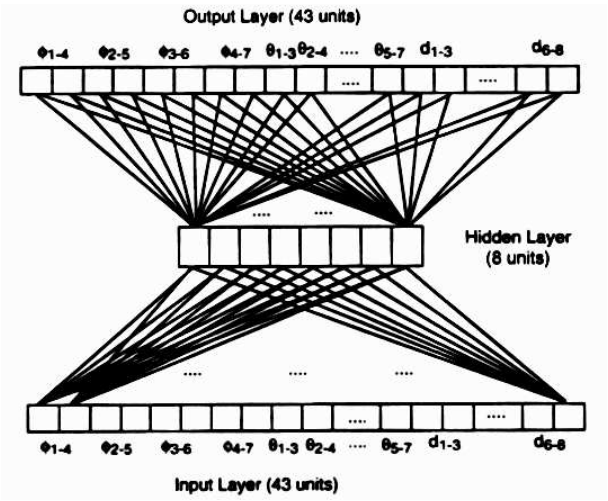


FIG. 2.14 – *Schéma représentant le réseau auto-associatif avec sa couche d'entrée à 43 neurones, la couche cachée de 8 neurones et la couches de sorties avec 43 neurones.*

Le groupe de Fetrow a utilisé 3 types d'informations structurales mises dans un réseau autoANN et ils se sont servis de la couche cachée comme nouveau descripteur des informations. Ils ont appliqué sur les données "compressées" par la couche cachée un algorithme classique de classification, les k-moyennes (ou *k-means*) qui est décrit dans l'Annexe 1.

Pour décrire les protéines, ils ont travaillé sur des fragments de 7 résidus consécutifs, qu'ils ont décrits par :

(-) 15 distances liant les C_α (sauf pour les C_α contigus). Ayant observé que la distribution des distances est bi-modale, chaque distance a été codée sur deux valeurs (bits). La valeur entre les deux pics de la courbe bi-modale a été prise comme seuil (τ). Si la distance d_i est inférieure à τ, le second bit est mis à zéro, et, dans le premier le rapport d_i/τ est introduit. Si la distance d_i est supérieure à τ, le premier bit est mis à un, et, dans le second le rapport $d_i - \tau/v_m - \tau$ est introduit. v_m est la valeur maximale des distances

observées. Ce système donne pour 15 distances 30 valeurs, toutes normalisées entre 0 et 1.

(-) les 4 angles dièdres, qu'ils ont codé avec leurs sinus et cosinus, soit 8 valeurs au final, ce qui élimine les problèmes de rotation.

(-) 5 angles de valences ($C\alpha_i$,$C\alpha_{i+1}$,$C\alpha_{i+2}$,$C\alpha_{i+3}$).

Ainsi 7 résidus sont traduits en un vecteur de 43 composantes, toutes comprises entre zéro et un. Pour le premier autoANN [208, 209], la moitié de leur base de données (6 422 fragments) a été utilisée pour l'apprentissage, le reste (6 242) servant pour sa validation. Ensuite, les auteurs utilisent les vecteurs recodés par la couche cachée (8 valeurs au lieu de 43) et les séparent en 6 groupes grâce à la méthode des k-moyennes ou *k-means* (cf. Annexe 1).

Leur première étude [208] (il n'y a aucune évolution notable dans leur seconde publication [209]), montre surtout une recherche de la corrélation entre leurs 6 blocs obtenus et les structures secondaires. Ainsi, plusieurs initialisations ont été testées et le maximum de ressemblance entre les résultats dans les différentes expériences a été recherché, les blocs obtenus devant être les plus proches possibles.

Différentes remarques sont à faire sur cette étude. Aucune notion de ressemblance 3D n'est mentionnée. Les auteurs décident de conserver 6 groupes soit 6 blocs. Ces blocs ne sont décrits que comme une entrée en hélice, une partie centrale, une extrémité C-terminale. Il en va de même pour les feuillets. La répartition précise dans les différents blocs des boucles (près de 50% de la base de données) n'est pas explicitée. Avec une absence totale de vérification de la fiabilité structurale et de la variabilité des blocs, des questions restent en suspens. Dans la même perspective, il aurait été intéressant de savoir ce que donne directement une méthode de classification sur les observations recodées en 43 composantes.

Fetrow et collaborateurs ont repris ce travail en 1997 avec une nouvelle base et aboutissent à des résultats similaires. Un travail intéressant est fait concernant les successions de doublets de blocs (*i,i+1*) et des triplets (*i,i+1,i+2*).

En conclusion, aucune précision n'est donnée quant à l'approximation structurale qu'engendrent les blocs. Cette information n'est retrouvée indirectement que dans la base de données mise sur le web *ftp://ftp.cs.albany.edu/pub/compbio/db/2.2A/cat/*, et, l'absence d'explication quant aux chiffres donnés la rend difficilement utilisable. Les doublets et triplets de blocs ne

sont donnés qu'en fonction de leurs fréquences attendues et observées. Une étude statistique plus élaborée aurait permis de voir les plus significatifs (qui ne sont pas obligatoirement les plus fréquents). Les quelques exemples montrant l'intérêt de la méthode sont peu nombreux et reposent principalement sur des successions très rares (entre 1 et 4 observations dans la base de données).

Plus récemment, Fetrow et Berg [52] ont tenté de mettre en évidence une relation existant entre leurs blocs et les chaînes latérales. Ils ne présentent que quelques graphiques sans véritable conclusion.

2.3.3.3 Utilisation des Chaînes de Markov Cachées (Camproux *et al.*, 1999)

Le but de cette étude est de définir un ensemble de prototypes pour approximer la structure tridimensionnelle des protéines en tenant compte principalement des transitions entre ces prototypes. Ils sont appelés blocs structuraux pour la reconstruction (Structural Building Blocs ou SBBs) [23]. La base de données est composée de 100 protéines ayant moins de 25 % de similitude de séquences [84, 83], soit 19 317 résidus.

La méthode se base sur l'utilisation des Chaînes de Markov Cachées (CMC, ou Hidden Markov Model, HMM) [149, 22] et utilise comme descripteur de la structure des protéines des distances entre C_α et une projection dans un plan. Les fragments de la base de données ont été découpés en fragments chevauchants d'une longueur de 4 résidus. La figure 2.15 montre les 3 distances utilisées: distance d_1 entre le $C\alpha_i$ et le $C\alpha_{i+2}$, distance d_2 entre le $C\alpha_i$ et le $C\alpha_{i+3}$ et distance d_3 entre le $C\alpha_{i+1}$ et le $C\alpha_{i+3}$. Une quatrième distance est calculée. Il s'agit de la distance d_4 qui est une projection du $C\alpha_{i+3}$ dans le plan défini par les $C\alpha_i$, $C\alpha_{i+1}$ et $C\alpha_{i+2}$. Cette dernière distance est normalisée par le produit des deux premières distances. Elle permet de donner la direction du fragment. Si d_4 est proche de 0, le fragment est allongé, sans volume, sinon il décrit une certaine torsion.

Le modèle est complexe. Il tend à maximiser le passage d'un état i à un nombre d'états limités. Il demande l'établissement au préalable d'un type de loi. Ici, la dispersion des observations associées à chaque état (SBB) a été considérée comme gaussienne.

Après avoir testé plusieurs possibilités, Camproux et collaborateurs [23] ont décidé de conserver un jeu de 12 blocs protéiques.

L'Analyse en Composantes Principales (ACP) effectuée sur la base de données codée suivant

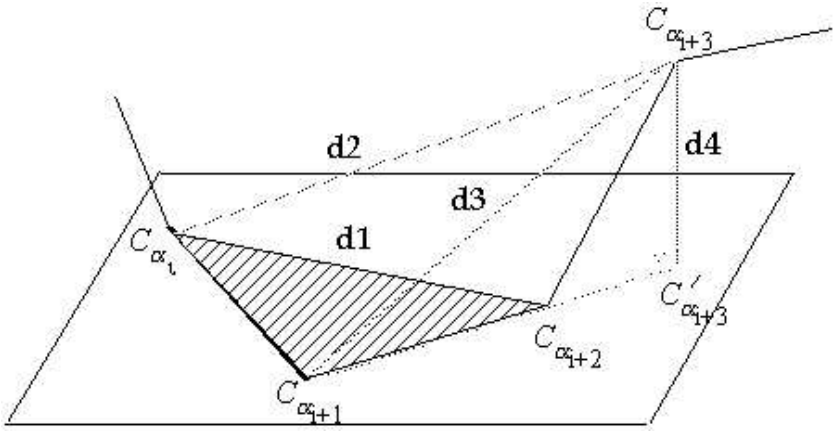


FIG. 2.15 – *Les 3 distances inter-C_α et la projection décrivant les fragments de longueurs 4, avec d_1 distance entre $C\alpha_i$ et $C\alpha_{i+2}$, d_2,$C\alpha_i$ et $C\alpha_{i+3}$ et d_3,$C\alpha_{i+1}$ et $C\alpha_{i+3}$, la distance d_4 est une projection du $C\alpha_{i+3}$ dans le plan $C\alpha_i$, $C\alpha_{i+1}$, $C\alpha_{i+2}$.*

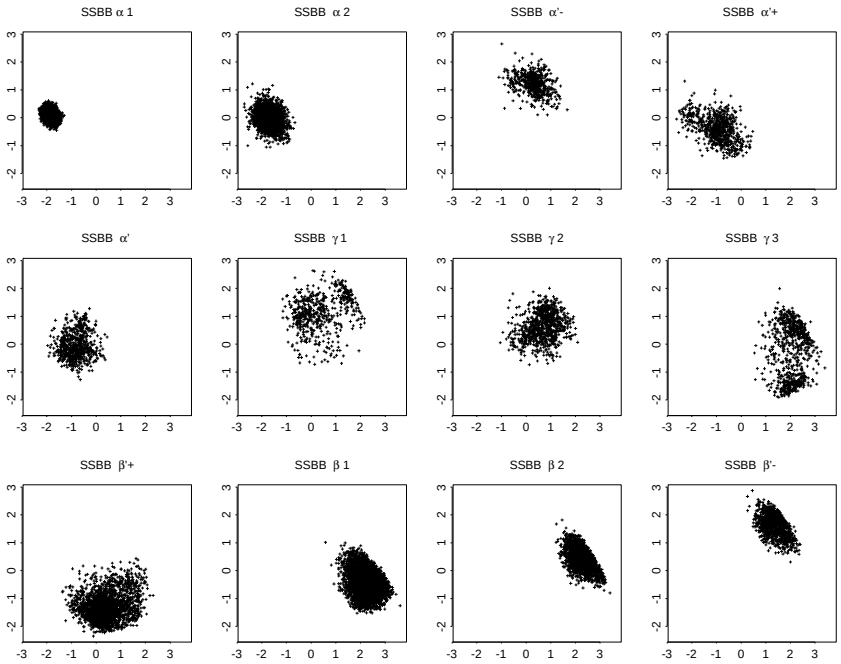


FIG. 2.16 – *Représentation des 12 SBBs suivant les deux premières composantes principales de l'ACP effectuée sur les 4 distances décrites (cf. figure 2.15).*

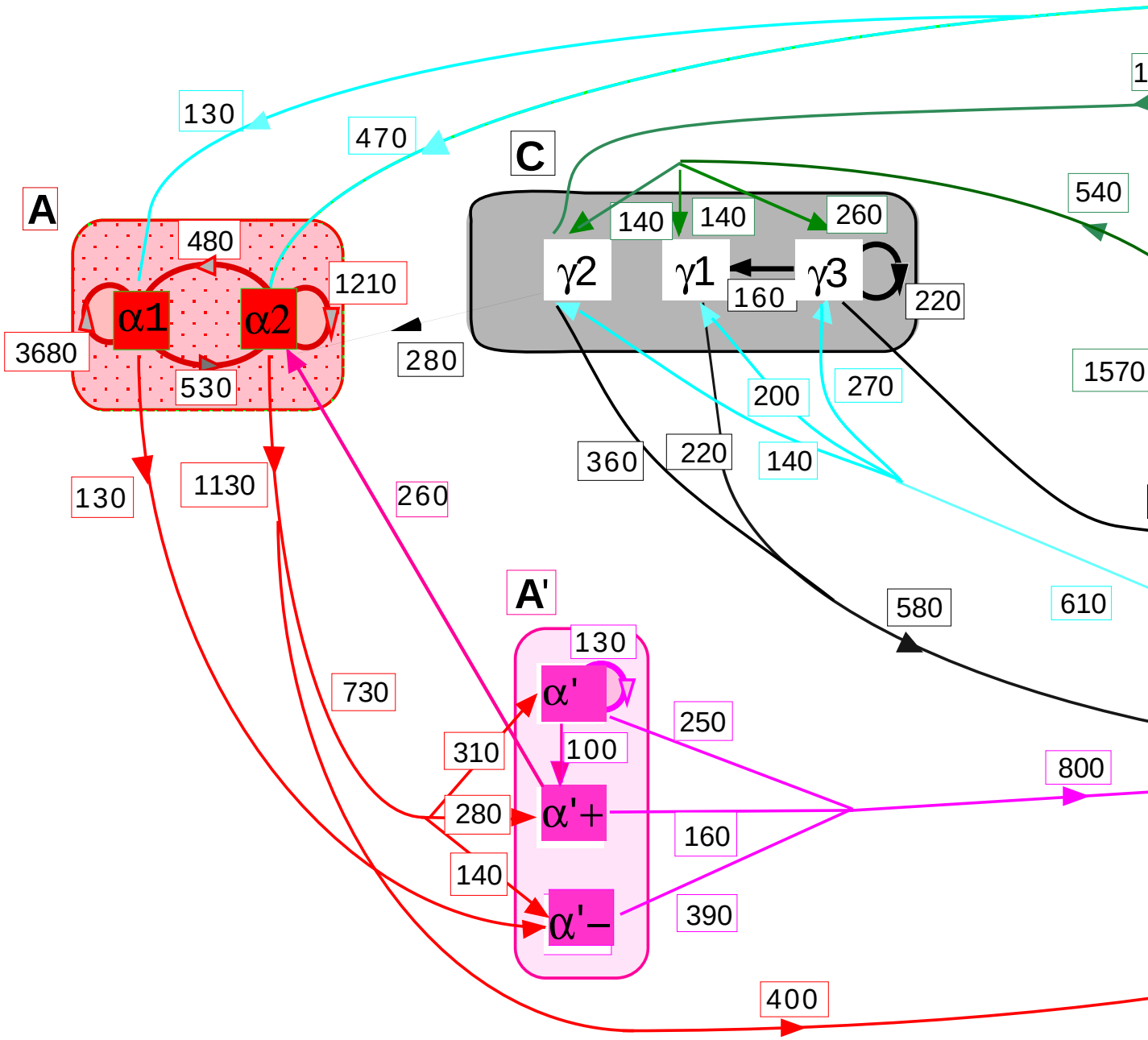


FIG. 2.17 – *Graphe de transition entre les 12 SBBs. Seules les occurrences supéri hélices α, en rosé, les SBBs Nterminaux des hélices α, en gris les SBBs impliqu bleu les SBBs aux extrêmités des feuillets β.*

les 4 distances montre certaines tendances distinctes suivant les types de structures secondaires. La figure 2.16 montre le nuage de point associé à chaque bloc obtenu. On voit clairement l'efficacité de la méthode pour bien discriminer les différents groupes qui sont tous bien compacts.

Deux SBBs sont caractéristiques des hélices α, SBB α_1 et SBB α_2, le premier représente 63,3 % des hélices α de la base de données, le second 28,3 %. Ils ne restent donc que 8,4 % des hélices α assignées à un autre bloc. De même, deux SBBs sont caractéristiques des feuillets β, SBB β_1 et SBB β_2. Le premier représente 57,6 % des feuillets β et le second 21,1 %. Il reste donc 21,6 % des feuillets associés à un autre bloc. Deux SBBs sont particulièrement liés aux boucles, les SBBs γ_1 et γ_2. Les autres sont liés à des entrées et/ou des sorties de structures répétitives. Les SBBs obtenus sont particulièrement stables d'un point de vue structural avec des *RMSd* compris entre 0,15 Å et 1,4 Å.

Un des intérêts des CMC est l'apprentissage basé sur les transitions. Ainsi la figure 2.17 montre le graphe des transitions existantes entre les 12 SBBs, pour des occurrences supérieures à 100.

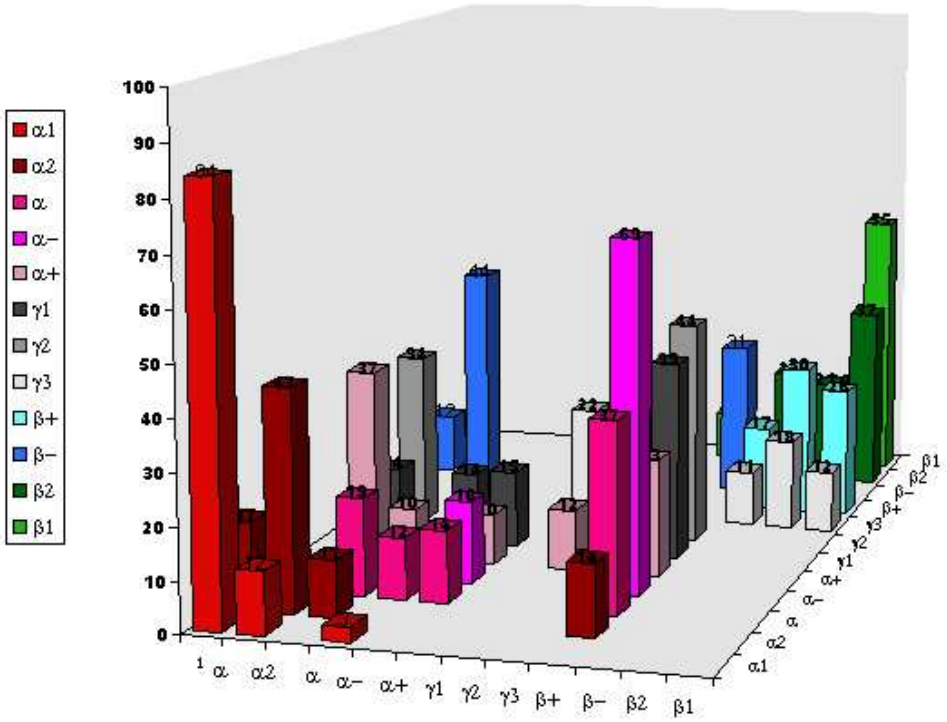


FIG. 2.18 – *Matrice de transition entre les 12 SBBs.*

La polarité existante entre les SBBs est extrêmement marquée. La matrice de transition représentée sur la figure 2.18 montre clairement le regroupement qui existe entre chaque groupe de SBBs qui suit bien les structures secondaires mais avec une certaine flexibilité. Ainsi le bloc β_+ (encore noté γ_β) possède 5 transitions possibles vers un autre type de SBB. Ce type d'apprentissage permet d'apprendre en suivant les transitions, mais sans focalisation excessive quand cela n'est plus possible.

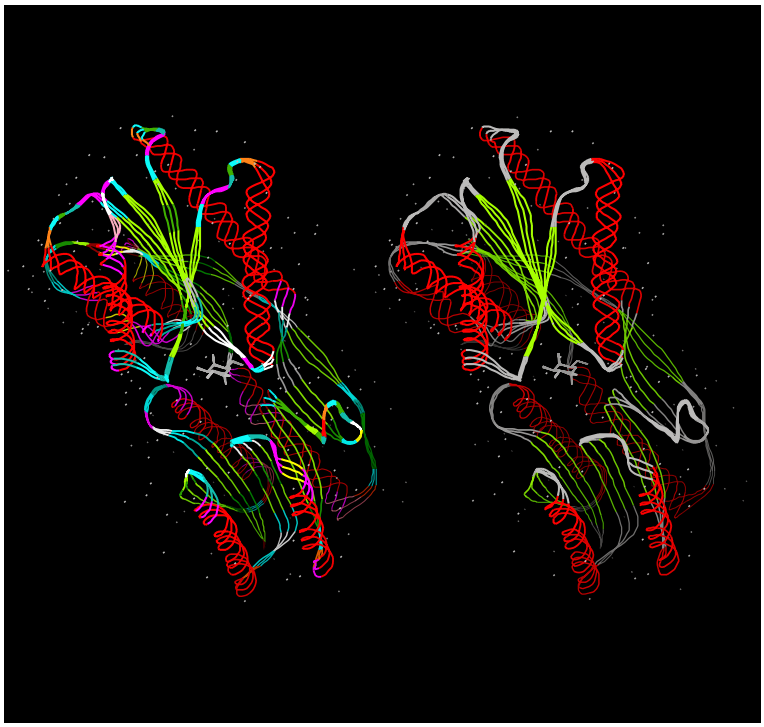


FIG. 2.19 – *Comparaison entre l'attribution par les 12 SBBs à gauche et les structures secondaires à droite avec l'aide du logiciel XmMol [194].*

La figure 2.19 montre la protéine de liaison au L-arabinose (code PDB: 8abp) colorisée à gauche suivant les SBBs, à droite suivant la définition classique des structures secondaires. On observe ainsi nettement la bonne correspondance entre les deux types d'information avec des changements ponctuels. Cette figure montre l'intérêt de ce type d'approche qui permet de bien distinguer, par exemple, le début d'une hélice assigné à un bloc α'_+ colorié en jaune.

Une information pertinente à prendre en compte est alors la distribution en acides aminés. Les distributions classiques associées aux structures répétitives sont retrouvées, comme

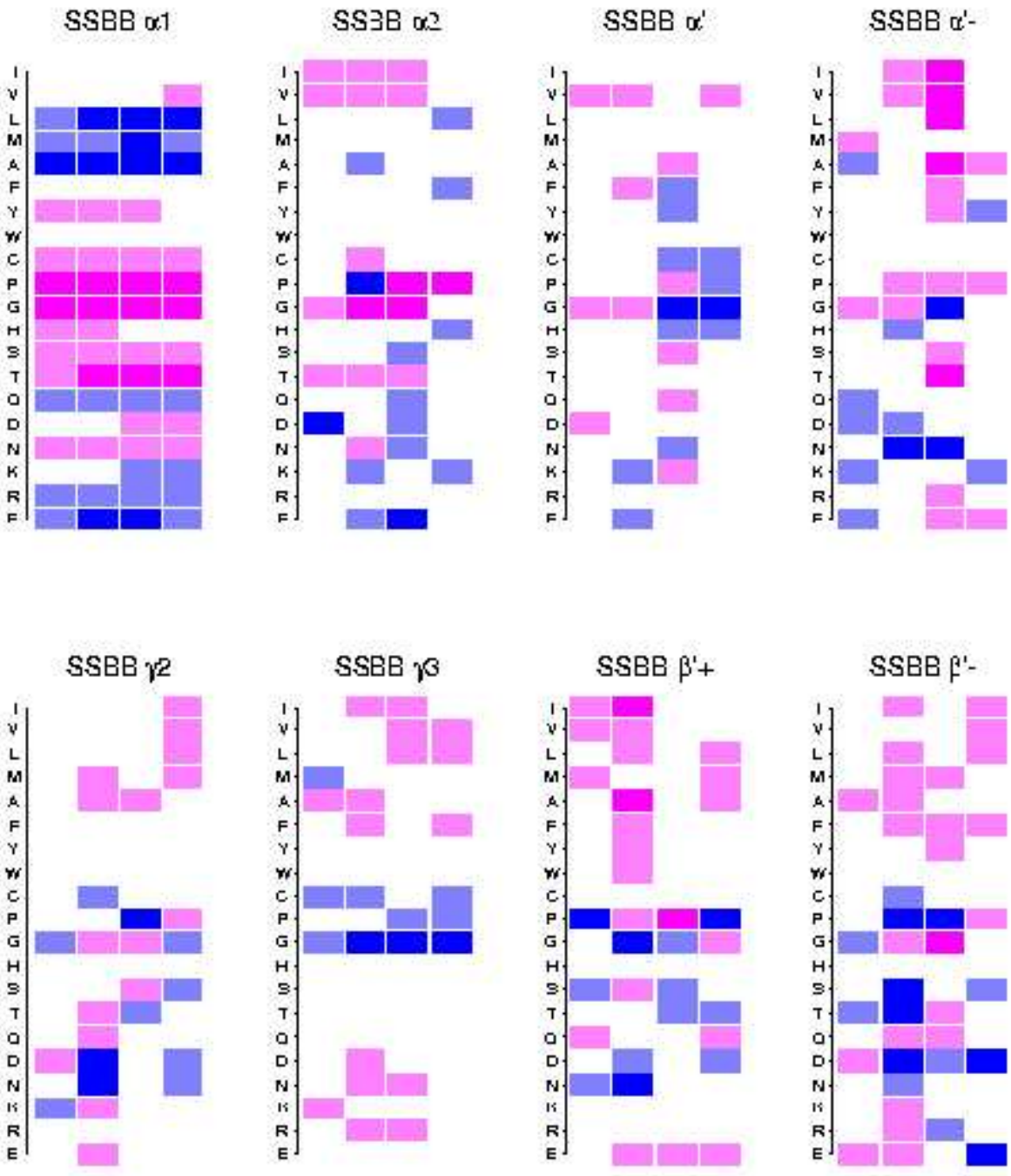


FIG. 2.20 – *Fréquences des acides aminés en chaque SBB pour les 4 positions*
représentation, en rosé sous-représentation). Les couleurs pâles représentent des
les autres sont supérieures à 4. Pour une définition du Z-score cf. paragraphe 3.

des acides aminés hydrophobes associés au bloc α_1. Le SBB α_2 du fait de sa composition en Acide Aspartique, Proline et Acide Glutamique semble être plutôt un coude α. Le seul point véritablement critique est la longueur des blocs. 4 C_α ne permet pas l'établissement de liaisons hydrogénes.

Deux travaux distincts ont suivi l'élaboration de cet alphabet structural, tout deux portant sur une analyse des parties boucles entre deux structures répétitives α et/ou β [24, 21].

2.3.4 Une méthode d'apprentissage sur la séquence et la structure (Bystroff et Baker, 1998)

Le travail de Bystroff et Baker concerne l'obtention d'un certain nombre de blocs structuraux protéiques (I-sites) types pour faire de la prédiction. Ils cherchent à optimiser la relation séquence-structure (1998, [19]). Les 471 protéines sont issues de la base de données HSSP [172], possèdent moins de 25% d'identités de séquences, et sont groupées en familles d'homologues.

Ce travail est à l'heure actuelle le plus abouti et celui qui a généré le plus grand nombre d'applications. Les I-sites débutent en réalité trois ans avant leur réalisation par un travail de Han et Baker. Ils mettent au point une méthode de regroupement de séquences ayant de fortes similitudes de séquences par une méthode proche des k-means [74]. Dans un second temps, ils mettent en relation le fait que les groupes qu'ils ont obtenus peuvent correspondre à un ou plusieurs types de repliements protéiques [75]. Ils se limitent à une simple description de répartitions en structures secondaires. Ils définissent alors directement à partir de leur groupes certaines structures plus ou moins précises [76].

La figure 2.21 récapitule l'ensemble du processus de création des I-sites. Les 471 protéines issues de HSSP sont alignées en une centaine de groupes par un alignement multiple. Certaines protéines sont exclues si leur résolution est trop mauvaise, s'il y a plusieurs ponts disulfures ou si la protéine est membranaire. Les longues boucles semblent avoir été exclues aussi.

Ensuite un regroupement de fragments est effectué à l'aide de la méthode des k-moyens (ou *k-means*, cf. Annexe 1) avec l'utilisation de profils sur les acides aminés. K groupes dont les longueurs sont variables sont obtenus.

Ces K groupes sont ensuite regroupés suivant leur similitude de repliement. Cette similitude

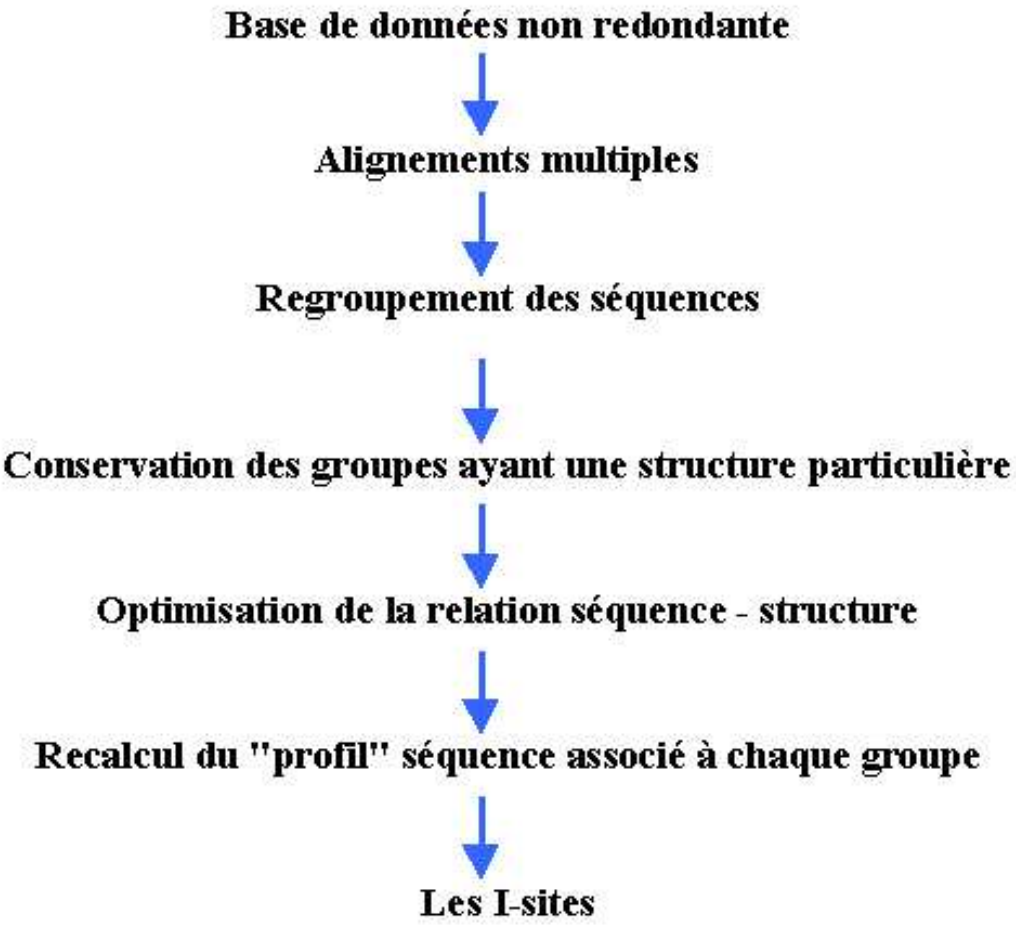


FIG. 2.21 – *Principe de la création des I-sites.*

est mesurée par deux critéres :

- dme :la "distance matrix error" qui revient au classique *RMSd* sur les C_α.
- dma :la déviation maximale angulaire : l'angle dièdre (ϕ ou ψ) ayant l'écart le plus important entre les deux structures :

$$dma(L) = max_{i=1,L-1}(\delta\phi_{i+1}, \delta\psi_i)$$

Pour créer un groupe, on sélectionne les "meilleures" séquences de chacun des K groupes. Ces séquences sont définies comme les plus proches du consensus (*centre*) de chaque groupe. Deux fragments protéiques sont considérés structurellement similaires si leur *dme* est inférieure à 1.4 Å et leur *dma* est inférieure à 120°. Cette méthode permet de regrouper leurs K groupes-séquences en K' groupes-structures. Une nouvelle optimisation est alors effectuée: elle consiste à ne conserver que les fragments les plus proches sur le plan de la séquence et à recalculer le profil séquence moyen. Les 400 séquences les plus proches servent de nouveau centre et le processus recommence. Il suffit de 3 à 5 cycles d'apprentissage pour stabiliser les groupes. Une autre étape d'optimisation suit, mais elle est peu décrite, et sert de dernier filtrage entre la séquence et la structure.

Les figures 7.7 à 7.11 montrent les $K' = 13$ groupes finaux obtenus à partir des $K = 82$ groupes obtenus sur les séquences. Les figures présentant les I-sites sont mises en Annexe 3. Les tailles varient au départ entre 3 et 17 résidus. Après le second regroupement, les longueurs moyennes observées ne sont pas données.

Les groupes ou I-sites sont fortement liés aux structures secondaires répétitives avec pour les hélices α 7 sites : le I-site 1, une hélice α amphipatique (cf. figure 7.7a), le I-site 2, hélice α non polaire (cf. figure 7.7b), le I-site 3, extrémité C-terminale d'hélice α Glycine dite de type 1 (cf. figure 7.7c), le I-site 4, extrémité C-terminale d'hélice α Glycine dite de type 2 (cf. figure 7.8a), le I-site 5, extrémité C-terminale d'hélice α Proline (cf. figure 7.8b), le I-site 6, hélice α mélée (cf. figure 7.8c), le I-site 7, extrémité N-terminale d'hélice α Sérine (cf. figure 7.9a), pour les feuillets β 2 sites : le I-site 8, feuillet β amphipatique (cf. figure 7.9b) le I-site 9, feuillet β hydrophobe (cf. figure 7.9c), pour les boucles 4 sites : le I-site 10, coude β Aspartate (cf. figure

7.10a), le I-site 11, épingle β Sérine (cf. figure 7.10b), le I-site 12, épingle type I étendu (cf. figure 7.10c), le I-site 13, coude type II divergeant (cf. figure 7.11).

Plusieurs points donnent matière à discussion sur cette première partie. Tout d'abord, l'importance des I-sites en relation avec les hélices α avec 3 I-sites pour la partie centrale, 1 pour les N-terminales et 3 pour les extrêmités C-terminales avec les deux types d'acides aminés classiques Proline et Glycine. Seuls deux I-sites sont clairement attachés aux feuillets β. Les 4 I-sites associés aux boucles représentent des changements brusques et courts. Ce dernier fait est peut-être dû à l'élimination des boucles les plus longues de l'apprentissage.

Le second point est l'absence de données concernant la structure des I-sites. Leur variabilité n'est pas clairement exprimée et ne concerne que le résultat de la prédiction.

La méthode se base sur les profils moyens obtenus précédement. Les scores obtenus sont pondérés en fonction de ceux obtenus dans la base d'apprentissage. Il convient de s'arréter particulièrement sur la méthode de calcul du taux de prédiction. Elle consiste à calculer le I-site le plus probable pour un fragment de 8 résidus consécutifs. Ensuite il convient de comparer les angles dièdres du fragment avec ceux du I-site et, si la *dma* est inférieure à 120^o, les 8 sites sont considérés comme correctement prédits :

$1 \leftarrow$ si $dma < 120^o$

$0 \leftarrow$ si $dma > 120^o$

Une utilisation conjointe de PHD [165] avec les I-sites est possible, principalement pour trouver des séquences homologues. L'évaluation par PHD n'est pas une simple prédiction par type de structures secondaires, mais une redéfinition des angles des I-sites selon la concordance entre la prédiction de PHD et des I-sites.

55 nouvelles protéines sont testées et un taux de prédiction global de 48% avec les I-sites est trouvé, et un taux de 54% en combinant leur méthode avec la prédiction des structures secondaires par PHD. Le tableau 2.8 montrent les différents taux de prédiction selon le type de protéines avec les I-sites seuls, avec la méthode PHD et les deux combinées.

Ainsi, Bystroff et Baker proposent un certain nombre de blocs structuraux (13) et font la prédiction des I-sites les plus probables à partir de la séquence [19]. Un point fort positif est l'existence d'un indice de confiance à 5 niveaux vis-à-vis de la prédiction. Toutefois, leur méthode de calcul de la prédiction est particulière et peut induire en erreur. Leur mesure fait

groupe protéique	nombre de protéines	I-sites	PHD	Combiné
tout α	8	40	55	55
tout β	6	51	32	51
$\alpha + \beta$	41	50	41	54
Total	55	48	43	54

TAB. 2.8 – *Prédiction sur la série de 55 protéines non utilisées durant l'apprentissage.*

qu'un fragment de 8 résidus est intégralement considéré comme bon, si chaque angle du I-site correspondant est à moins de 120^o de la réalité. En prenant le cas extrême d'un peptide de longueur 16, si le premier et le dernier I-site sont bons, le taux de prédiction est de 100%. Ceci même quand les positions suivantes sont intégralement incompatibles.

J'ai recalculé un pourcentage de prédiction modifié *P'* pour voir la sensibilité de leur métrique :

$$P' = \frac{(N_p - (N_z \times 7))}{N_t}$$

avec N_p le nombre de sites correctement prédits, N_z, le nombre de zones continues de bonne prédiction et N_t le nombre total de résidus dans la protéine. Le principe est fort simple, si tous les sites correctement prédits sont continus, une seule zone existe donc. Alors, N_z=1, et seuls 7 sites sur les bords doivent être soustraits. La figure 2.22 montre le résulat pour une protéine de 100 résidus de longueur. La descente est moins accentuée avec une protéine de taille supérieure, mais l'exemple me paraît frappant. Sur la figure, le dernier point hyp(1) représente l'hypothèse selon laquelle tous les I-sites correctement prédits sont indépendants et aucun chevauchement n'existe.

Le calibrage de ce type de fonction est complexe, car aucune idée du nombre de zones de I-sites correctement déterminées n'est accessible, en se réferant aux structures secondaires, il serait possible d'évaluer un nombre approximatif. Un calcul rapide montre un nombre de zones au minimum supérieur à 7.

Les résultats récapitulés dans le tableau 2.8 montrent que les hélices α malgré leur importance dans les I-sites sont moyennement prédites, mais que les feuillets β ont un taux supérieur à celui généralement observé.

Avant même la publication de l'article des I-sites [19], Baker et Bystroff les ont utilisés pour

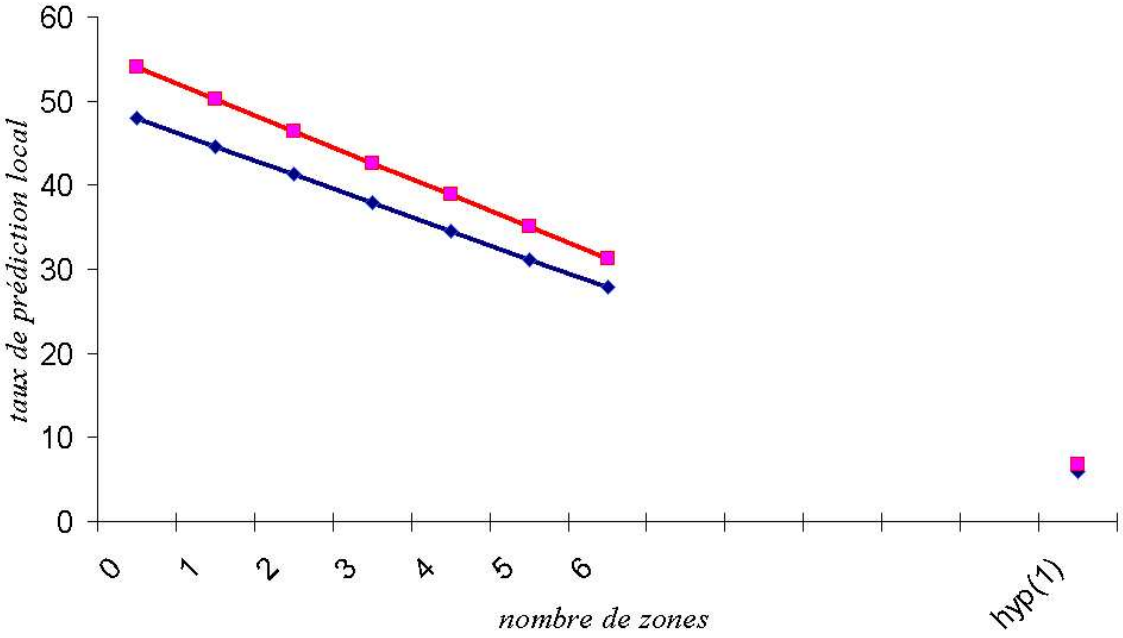


FIG. 2.22 – *Exemple du calcul du taux de prédiction pour une protéine ayant $N_t = 100$ résidus, avec en rouge les taux associés à la prédiction des I-sites combinés avec PHD et en bleu les I-sites seuls. hyp(1) représente le taux associé à l'hypothèse selon laquelle aucun site correctement prédit ne serait chevauchant avec un autre.*

prédire des structures protéiques lors de CASP 2 [18]. Sur les 8 séquences, seules 2 ont été traitées par les I-sites combinés avec PHD. Les autres n'ont utilisé que I-sites. Cela ne change pas les taux qui sont de 39% pour les protéines avec les I-sites et 41% avec la méthode combinée.

Ils appliquent, ensuite, leur méthode à la modélisation d'une zone dans un domaine SH_3 [206]. Les scores de confiance sont particulièrement élevés dans cette zone. De plus, des données RMN sur la structure corroborrent leur application. Cet exemple pose néanmoins un autre problème: l'influence de l'alignement. Leurs I-sites fort bien prédits sont les parties les plus conservées, d'un point de vue séquence.

Dans le même laboratoire, Simons et collaborateurs ont mis au point une méthode de modélisation *ab initio* sur un principe de statistique bayésienne avec recuit simulé [179]. Dans leur modèle bayésien le principe de relation séquence-structure vu par Han et Baker [75] est utilisé. Comme dans d'autres méthodes *ab initio*, seules les protéines de petites tailles ont un repliement acceptable. Ils essayent par la suite une série d'améliorations dont l'utilisation des I-sites, comme filtre au même titre que les structures secondaires, ce qui n'augmente pas l'efficacité de la méthode [180]. Pour CASP 3, l'utilisation de d'une nouvelle approche utilisant les I-sites et la méthode *ab initio*, travaillée cette fois-ci indépendament, a donné quelques très bons résultats

[178].

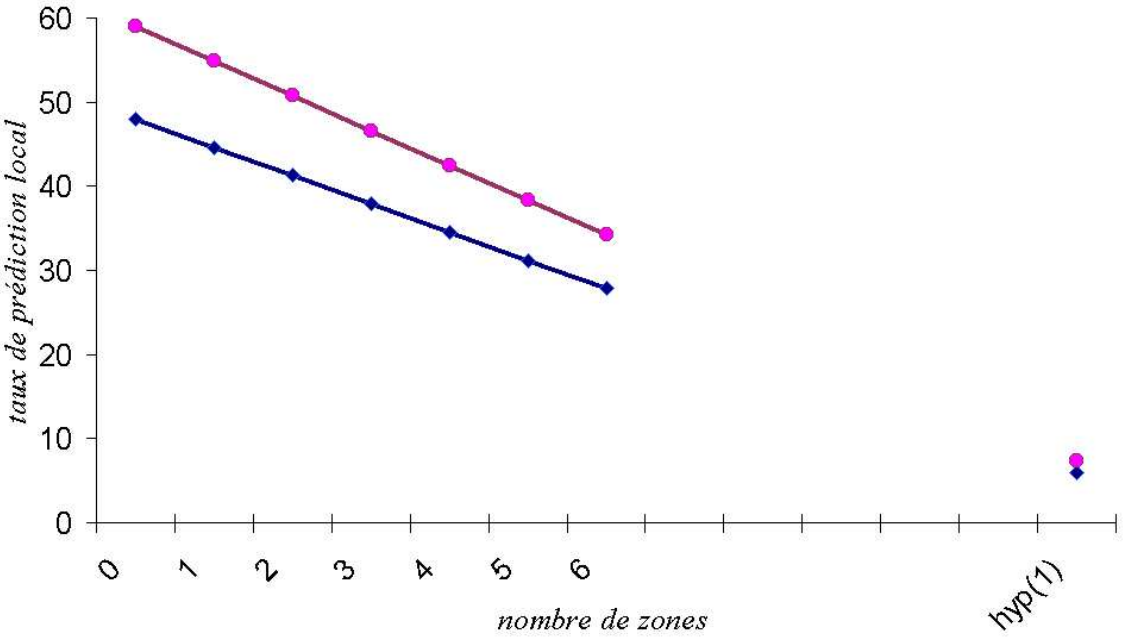


FIG. 2.23 – *Exemple du calcul du taux de prédiction pour une protéine ayant N_t = 100 résidus, avec en rouge les taux associés à la prédiction des I-sites utilisés dans HMMSTR [20] et en bleu les I-sites du papier original [19]. hyp(1) représente le taux associé à l'hypothèse selon laquelle aucun site correctement prédit ne serait chevauchant avec un autre.*

Enfin, Bystroff et collaborateurs ont utilisé les I-sites pour prédire par une méthode de Chaînes de Markov Cachées les structures secondaires (nommé HMMSTR) avec un taux final de 74,3% [20]. Pour ceci, ils ont utilisé les 13 I-sites [19], plus trois nouveaux et ensuite ont effectué leur recherche en utilisant les angles dièdres pour catégoriser le type de structure prédite. Un détail intéressant est le passage du taux de prédiction des I-sites de 48 à 59%, alors que le nombre de sites augmente, ce qui est surprenant. La figure 2.23 montre le calcul du taux de prédiction modifié P' dans ce nouveau cas. En réalité, avec quelques zones distinctes, le taux de prédiction est partiquement identique, ce qui peut expliquer ce taux "élevé" de prédiction.

2.3.4.1 Récapitulatif des alphabets structuraux et conclusion

Un petit résumé des différentes méthodes et techniques utilisées montrent bien l'avancée réalisée en une dizaine d'années. Elle est présentée dans le tableau 2.9. Commençant avec un nombre très limité de protéines, les approches sont devenues plus complexes et plus sûres, utilisant un nombre conséquent de résidus. Toutefois, elles montrent aussi l'hétérogénéité des techniques et par la suite nous verrons les difficultés que cela engendre quand on désire comparer

Equipe	année	Nombre de protéines	Nombre d'acides aminés	Méthodes d'apprentissage Méthodes d'apprentissage	
Unger *et al.*	1989	4 / 82	426 / 12 973	k-means	
Rooman *et al.*	1990	75	12 978	regroupement hiérarchique	
Prestrelski *et al.*	1992	14	2 437	fonction	
Zhang *et al.*	1993	74	13114	autoANN	
Schuchhardt *et al.*	1996	136	24239	carte de Kohonen	
Fetrow *et al.*	1997	116	23 355	autoANN	
Bystroff et Baker	1998	471	*(NP)*	k-means	p
Camproux *et al.*	1999	100	19317	CMC	

TAB. 2.9 – *Récapitulatif des différents travaux, avec le nom de léquipe, l'année d utilisées et le nombre de résidus correspondants, la méthode d'apprentissage et le et leurs longueurs. (NP : non précisé).*

entre eux ces différents alphabets.

2.4 Conclusion

Les différentes méthodes d'analyse de la structure protéique montrent l'intérêt des structures secondaires, alphabet structural simple. Les progrès dans leur prédiction sont plus que prometteurs. Toutefois, même en supposant qu'ils puissent être parfaits, ils ne sont qu'un premier pas dans l'établissement d'une structure protéique 3D. En plus de l'analyse classique des boucles, les alphabets structuraux sont un outil précieux d'analyse pour comprendre l'architecture protéique et la répartition des acides aminés.

Les alphabets structuraux récents montrent de constants progrès et leur plus grande finesse, liée à la fois à un contrôle du nombre de blocs structuraux en jeu et à des méthodes d'obtention plus complexe. Toutefois des améliorations restent à faire, pour aboutir à un juste compromis entre une méthode d'obtention de prototypes structuraux simples utilisables dans une méthode de prédiction efficace qui permettent, en outre, une compréhension des différents acides aminés impliqués.

Chapitre 3

Apprentissage de la structure locale du squelette protéique

3.1 Objectif

L'objectif principal de cette thèse est l'élaboration et l'utilisation d'un alphabet structural dans une méthode de prédiction à partir de la séquence. Dans cette première partie de ma thèse, je développerai donc la méthode d'apprentissage des Blocs Protéiques (BPs). Ils permettent une description statistiquement représentative des conformations d'un ensemble de résidus consécutifs. Différent aspects de leur validation sur un plan structural seront détaillés. Dans le chapitre suivant, je développerai l'approche mise en place pour la prédiction locale de ces blocs à partir de la séquence.

Pour résumer ce chapitre, une nouvelle méthodologie a été mise au point pour obtenir un alphabet structural protéique. Elle se base sur les angles dièdres du squelette polypeptidique et procède par un apprentissage en deux étapes; une première non supervisée et une seconde qui tient compte des transitions qui existent entre blocs protéiques, d'une manière similaire à la technique des Chaînes de Markov Cachées (CMC) [149]. Ces blocs ont été conçus pour assurer une bonne approximation de la structure protéique. J'ai donc testé différentes séries ayant des nombres variables de blocs protéiques. La série qui a été conservée est celle dont le nombre de blocs protéiques répond au mieux à l'approximation de la structure protéique et à un taux de prédiction à partir de la séquence correct.

3.2 Méthode d'apprentissage

La méthode d'apprentissage que nous avons utilisée possède une certaine similitude avec celle de Schuchhardt et collaborateurs [175]. Il s'agit d'une méthode d'apprentissage proche des cartes auto-organisées de Kohonen [109].

Seront développés ici le type d'information prise en compte, avec la base de données et le codage des protéines en fonction des angles dièdres, la fonction de similitude entre deux fragments (pour savoir s'ils sont proches ou non), les deux étapes de la méthode d'apprentissage (non supervisée puis tenant compte des transitions) et enfin les critères permettant l'obtention d'un nombre différent de blocs protéiques pour chaque série.

3.2.1 Base de données

La base de données protéique est un critère primordial pour comprendre la relation existant entre la structure tridimensionnelle et la séquence (cf paragraphe 2.3, *Les alphabets structuraux*). Actuellement, la base de données des structures protéiques, la PDB (Protein Data Bank [6]) est composée de plus de 10 000 protéines. Conserver l'ensemble des protéines de la PDB reviendrait à travailler principalement sur quelques grands groupes, les plus "aisément" cristallisables (les plus nombreux dans la base alors), comme les lyzozymes, ce qui biaiserait les analyses (en particulier statistique). Ne conserver que des protéines ayant une faible homologie de séquence permet de limiter ce type d'erreur. La base utilisée est composée de 342 protéines ayant moins de 25% d'identité de séquence [84, 83].

Les structures secondaires ont été assignées au préalable par une méthode consensuelle ([31] et cf. paragraphe 2.1.2.1) de trois algorithmes DSSP [100], P-CURVE [181] et DEFINE [154]. Ensuite, les protéines ont été classées suivant la nomenclature de Michie et collaborateurs [127] qui considèrent 4 classes de protéines : tout α, tout β, α/β et non-classées.

3.2.2 Description des protéines

Nous avons choisi d'utiliser comme descripteur de la structure des protéines les angles dièdres ϕ et ψ. (cf. figure 3.1). La définition des angles dièdres est donnée au paragraphe 2.1.2.

L'angle ω n'a pas été pris en compte, car il reste particulièrement constant autour de 180^o. Sur la PDB de juillet 2000, moins de 0,5 % des résidus étaient en dehors de l'intervalle classique

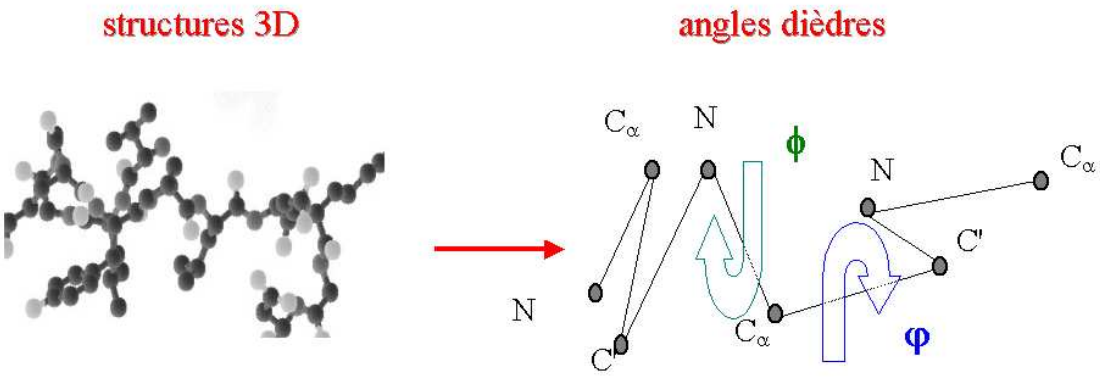


FIG. 3.1 – *Schéma représentant le calcul des angles ϕ et ψ à partir de la structure tridimensionnelle.*

[180°-10°;180°+10°]. En outre, l'angle ω suit directement les variations des angles ϕ et ψ [125]. Selon la zone de Ramachandran [ϕ,ψ], l'angle ω est statistiquement plus présent dans certaines zones. Leur utilisation aurait tendance à surestimer ce paramètre, 99,5 % ne jouant alors aucun rôle, les 0,5 % restant deviendraient trop important au regard de leur faible nombre.

La conformation des protéines a donc été décrite via leurs angles ϕ et ψ. Une protéine de longueur L est ainsi traduite en un vecteur de longueur 2 x (L-2) (le premier ϕ et le dernier ψ ne pouvant pas être calculés). L'intérêt d'utiliser des valeurs réelles des angles permet de ne pas avoir d'*a priori* particulier envers certaines zones préférentielles du diagramme de Ramachandran comme les zones répétitives et ainsi d'obtenir des blocs pertinents particulièrement au niveau des boucles.

3.2.3 Les fragments protéiques

Les 342 protéines de la base de données ont été ensuite découpées en fragment d'une longueur de $M = 5$ résidus consécutifs. Cette taille est suffisante pour décrire les courtes hélices qui ont une taille de 4 résidus [113] ainsi que les petits feuillets de 3 acides aminés [31]. De plus, 5 résidus est une taille acceptable pour permettre l'établissement de liaisons hydrogènes qui sont importantes d'un point de vue structural (cf paragraphe 2.1.2.1). Il faut bien noter que les fragments sont chevauchants; ainsi une protéine de longueur L sera représentée par L-4 fragments. La base de données comprend 86 628 fragments protéiques. Chaque fragment centré sur le carbone alpha ($C\alpha_n$) représente donc les carbones alpha $C\alpha_{n-2}, C\alpha_{n-1}, C\alpha_n, C\alpha_{n+1}$ et $C\alpha_{n+2}$. Lors de l'apprentissage, il sera représenté par un vecteur **V** de *2(M-1)* résidus, soit 8 angles

dièdres ($\psi_{n-2}, \phi_{n-1}, \psi_{n-1}, \phi_n, \psi_n, \phi_{n+1}, \psi_{n+1}, \phi_{n+2}$). Ainsi, un fragment (position i) possède 4 C_α en commun avec le fragment suivant (position $i+1$), soit 6 angles dièdres en commun.

3.2.4 L'apprentissage

Les protéines ont été séparées en deux groupes distincts. Le premier, utilisé pour l'apprentissage, comprend 228 protéines, ce qui représente 2/3 de la base de données. Le second comprend les 114 protéines restantes qui serviront pour la validation finale. Le but de l'apprentissage est de définir une série de Blocs Protéiques (BPs) qui représentent au mieux les fragments protéiques de la base de données. Aussi, chaque bloc protéique a une longueur identique au fragment soit *2(M-1)* valeurs, 8 angles dièdres.

3.2.4.1 Principe de l'apprentisage

La méthode utilise le principe des cartes auto-organisées de Kohonen (Kohonen Maps, ou Self-Organized Maps noté SOM [107, 109], cf. Annexe 1 et paragraphe 2.3.2.3). Elle procède par une lecture un certain nombre de fois (étape appelée cycle) de la base de donnée d'apprentissage. Dans la terminologie des SOMs, les neurones sont des classes d'objets. Leurs poids sont des informations moyennes associées à chaque neurone. Dans notre cas, il s'agit des blocs protéiques (*les neurones*) qui sont des vecteurs (*8 angles dièdres*).

Deux étapes d'apprentissage sont effectuées : la première consiste à apprendre les blocs protéiques seuls (par série de 5 C_α); la seconde permet de renforcer les transitions entre les blocs comme dans les Chaînes de Markov Cachées (CMC, [149]). Cette seconde étape est effectuée pour tenir compte de l'architecture des protéines. Elle revient à optimiser la succession des blocs protéiques. Pour expliciter ceci d'un point de vue structural, cette méthode revient à faire qu'une partie centrale, régulière, d'une hélice α assignée donc à un bloc "partie centrale α" aille préférentiellement vers un bloc "sortie d'hélice α". Cette optique permet, en outre, de n'avoir qu'un nombre limité de blocs "sorties d'hélices α". D'un point de vue plus formel quand un fragment protéique (*exemple*) en position i dans une protéine est associé à un bloc protéique (*neurone*) j, le fragment (*exemple*) en position suivante $i+1$ devrait être associé à un bloc protéique (*neurone*) k qui n'est pas n'importe lequel des B blocs (*neurones*), mais un des quelques v blocs (*neurones*) qui suivent normalement le bloc (*neurone*) j. Cette approche

ne repose sur aucune hypothèse ou loi de distribution *a priori*, contrairement aux CMC.

3.2.4.2 Mesure de similitude

La première mesure à effectuer est de savoir si deux fragments, ou un fragment et un bloc, sont proches structuralement. Pour calculer cette similitude, j'ai utilisé la différence moyenne entre les valeurs angulaires (ou *RMSda* pour root mean square deviations on angular values [175]). Il s'agit d'une distance euclidienne entre deux vecteurs $\mathbf{V}_1$ and $\mathbf{V}_2$:

$$RMSda(\mathbf{V}_1, \mathbf{V}_2) = \sqrt{\frac{\sum_{i=1}^{i=M-1}[\psi_i(\mathbf{V}_1) - \psi_i(\mathbf{V}_2)]^2 + [\phi_{i+1}(\mathbf{V}_1) - \phi_{i+1}(\mathbf{V}_2)]^2}{2(M-1)}}$$

avec $\{\phi_i(\mathbf{V}_1), \psi_{i+1}(\mathbf{V}_1)\}(resp.\ \phi_i(\mathbf{V}_2), \psi_{i+1}(\mathbf{V}_2))$ les séries de (2M-1) angles dièdres pour $\mathbf{V}_1(resp.\ \mathbf{V}_2)$. Les différences entre les angles sont calculées modulo 360^o.

Ainsi, durant la phase d'apprentissage, cette mesure est utilisée pour calculer la similitude entre un fragment et les différents blocs protéiques. La distance minimale permet d'associer un fragment à un bloc protéique.

3.2.4.3 Initialisation

Chaque bloc protéique PB_k, ou neurone, est initialisé par un vecteur $\mathbf{W}(k)$ constitué de 8 angles dièdres choisis soit dans les zones denses du diagramme de Ramachandran (cf. figure 2.4b), soit tirés aléatoirement dans la base de données. Le nombre initial de blocs B a été fixé arbitrairement.

3.2.4.4 Apprentissage non supervisé

A chaque cycle, les fragments sont tirés aléatoirement. La figure 3.2 récapitule l'ensemble des deux étapes du processus d'apprentissage, la figure 3.3 pour la première et la figure 3.4 pour la seconde.

Dans un premier temps, l'apprentissage ne tient pas compte des transitions. (1) Un fragment est choisi aléatoirement dans la base de données d'apprentissage. (2) Son vecteur d'observation $\mathbf{V}(m)$ est comparé à chacun des blocs protéiques PB_k à l'aide du critère du *RMSda*. B comparaisons sont donc calculées. (3) Le *RMSda* minimal correspond ainsi au bloc le plus proche

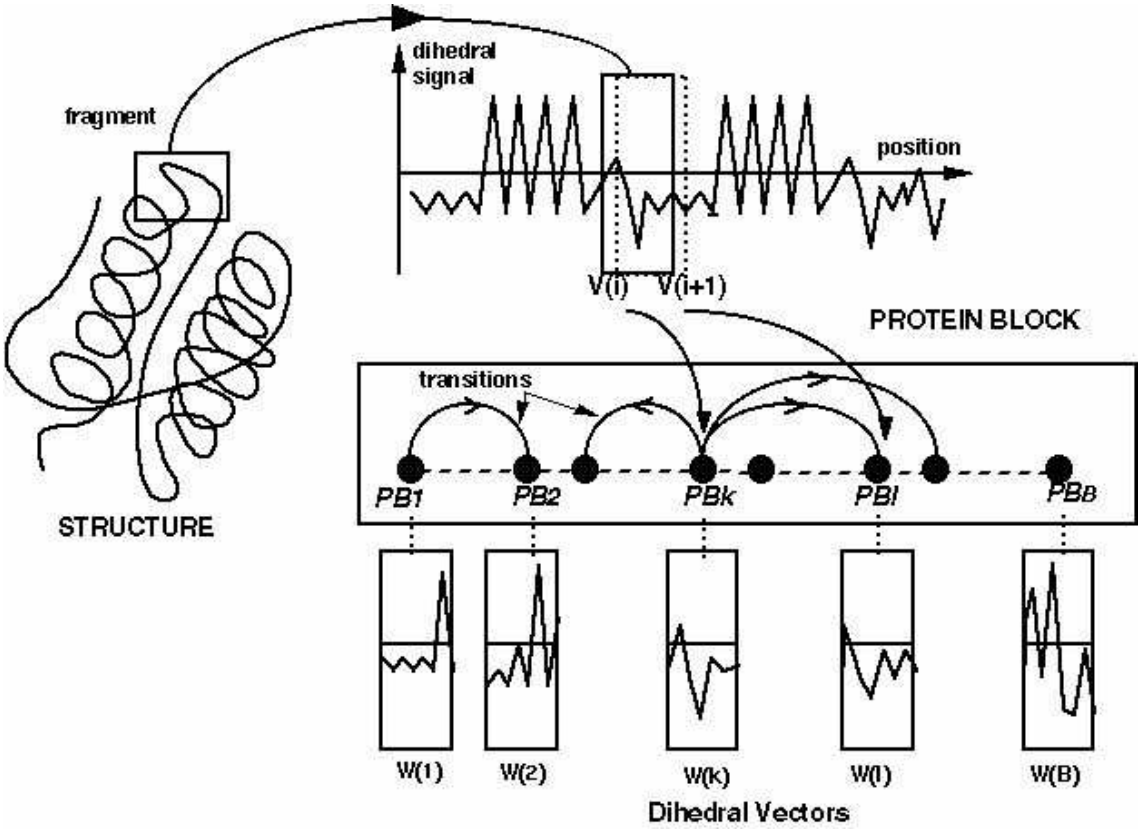


FIG. 3.2 – *Schéma global sur la méthode d'apprentissage des blocs protéiques. En partant de la protéine, à gauche, sont représentés, la traduction de la structure protéique en angles dièdres, le tirage aléatoire d'un fragment protéique, puis le choix du bloc le plus proche structuralement, et enfin la deuxième phase où l'on recherche parmis les blocs ayant un* RMSda *faible, celui qui a le meilleur taux de transition par rapport au précédent fragment.*

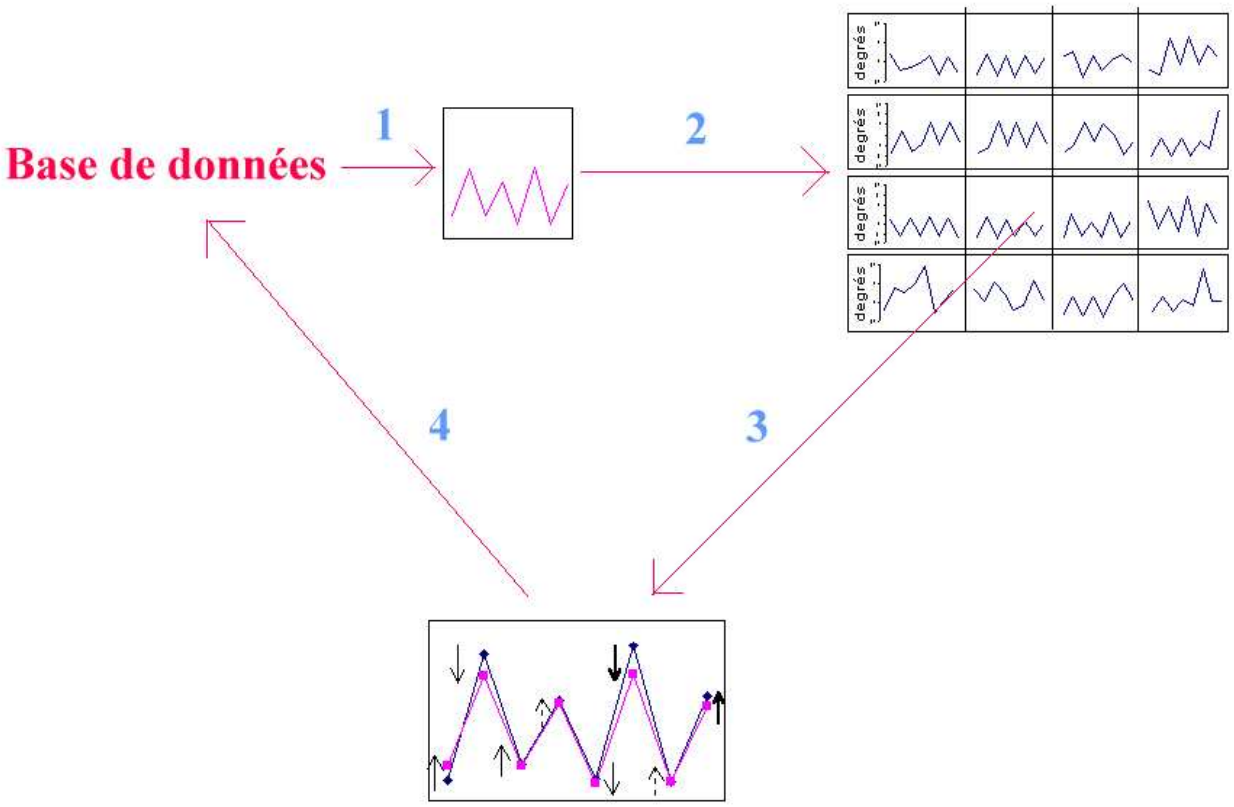


FIG. 3.3 – *Première étape d'apprentissage. (1) Un fragment est choisi aléatoirement dans la base de données protéique. (2). Les* B *distances sont calculées avec les* B *blocs existants. (3). Le bloc le plus proche du fragment tiré aléatoirement est modifé légèrement pour lui ressembler. (4). Le processus recommence au (1).*

structuralement du fragment. (4) Le vecteur $\mathbf{W}(k)$ du bloc le plus proche est très faiblement modifié pour ressembler à celui présenté au réseau $\mathbf{V}(m)$:

$$\mathbf{W}(k) \leftarrow \mathbf{W}(k) + (\mathbf{V}(m) - \mathbf{W}(k)).\nu(c)$$

avec $\nu(c)$ le coefficient d'apprentissage. Le symbole $\leftarrow$ veut dire "la valeur de gauche est remplacée par celle calculée à droite". Ce coefficient, initialement pris faible, i.e. ν_0 =0.02, décroit pendant l'apprentissage. $\nu(c)$ se modifie avec le nombre c de vecteurs d'observations déjà vus :

$$\nu(c) = \frac{\nu_0}{1 + \tau.c}$$

avec τ fixé arbitrairement à *1/N*, N représentant le nombre de vecteurs d'observations présents dans la base de donnée d'apprentissage. Dans notre cas, $N = 56\,442$. Ainsi, $\nu(c)$ est divisé par deux après le premier passage complet de la base.

L'apprentissage est itératif, un certain nombre de cycles C (= 15) est nécessaire pour définir des vecteurs optimaux $\mathbf{W}$ associés aux blocs. Chaque fragment n'est utilisé qu'une seule fois par cycle.

3.2.4.5 Apprentissage tenant compte des transitions

Après C cycles d'apprentissage, une première série de blocs protéiques a été obtenue (cf. figure 3.4). Ces derniers sont utilisés pour recoder en terme de blocs protéiques l'ensemble de la base de données d'apprentissage. Ceci permet de calculer une matrice de transition entre les BPs en comptant les occurrences des paires de BPs consécutifs et en les transformant ensuite en fréquences.

Le principe est un peu similaire au précédent. (1) Une protéine est tirée au hasard. (2) En première position, la recherche du BP_r, le bloc réel est effectué à l'aide du critère du *RMSda* minimal, comme précédemment. Pour les autres fragments, le processus est un peu différent. Leur vecteur d'observation $\mathbf{V}(m)$ est comparé à chacun des blocs protéiques PB_k à l'aide du critère du *RMSda* et B comparaisons sont donc calculées. Toutefois, le bloc vainqueur est choisi suivant des critères plus différents. (i) n blocs ayant un *RMSda* faible (inférieur à une valeur

donnée) sont sélectionnés. (ii) Le bloc protéique choisi est celui qui possède la fréquence de transition maximale entre le BP_r et BP_{r+1} pour les n blocs conservés. (4) Le processus est répété jusqu'au fragment C-terminal de la protéine. Ainsi, les transitions sont renforçées entre les différents blocs. À chaque cycle, la matrice de transition est réevaluée.

Au bout de C cycles d'apprentissage non supervisé, puis de C cycles d'apprentissage tenant compte des transitions, de nouveau C cycles d'apprentissage non supervisé sont effectués pour ne pas biaiser l'apprentissage.

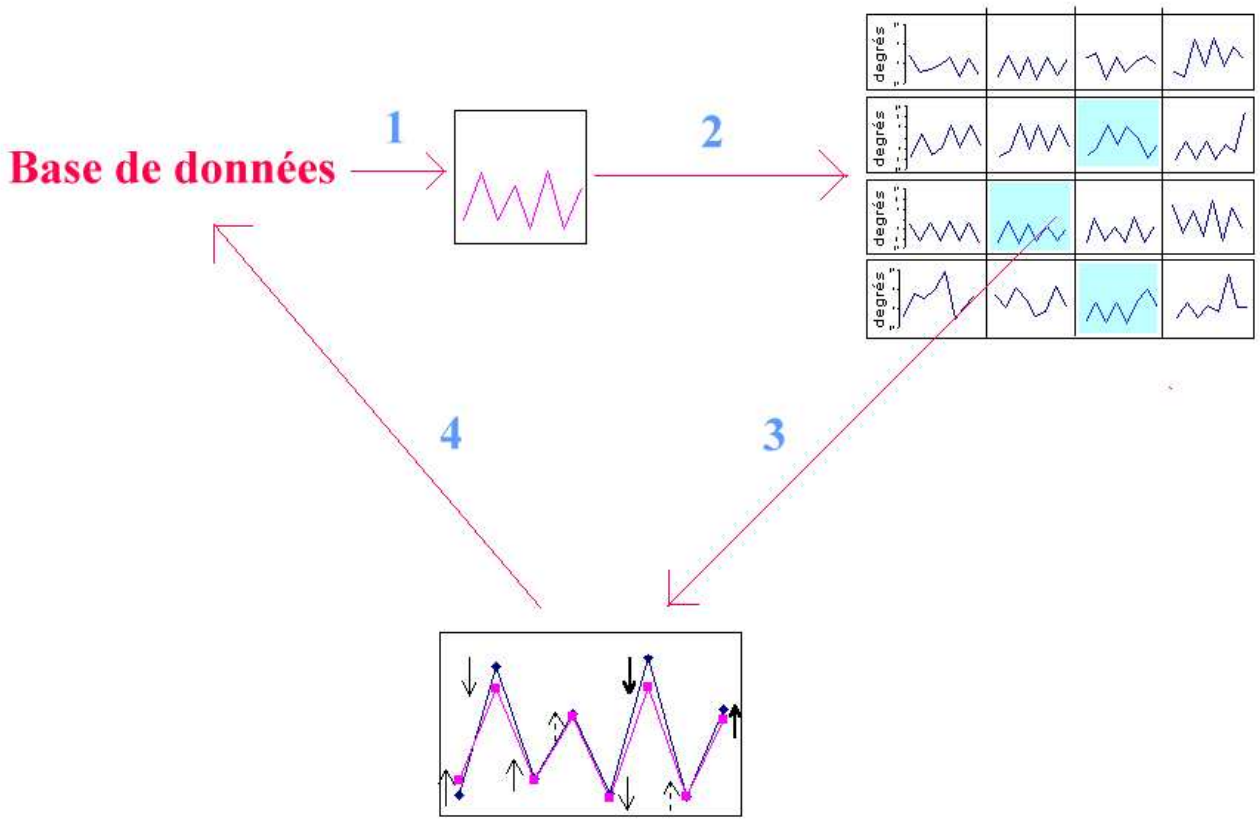


FIG. 3.4 – *Seconde étape d'apprentissage. (1) Un fragment est choisi aléatoirement dans la base de données protéique. (2). Les* B *distances sont calculées avec les* B *blocs existants.* n *blocs ayant un* RMSda *faible sont sélectionnés (les blocs mis en bleu clair). (3). Le bloc optimal est choisi parmi les* n *blocs, il représente la fréquence maximale de transition par rapport au bloc choisi précédemment (cf. le texte). Il est alors est modifé légèrement pour ressembler au fragment.(4). Le processus recommence au (1).*

3.2.4.6 Procédure de réduction du nombre de blocs protéiques

Partant d'un nombre particulièrement important de B (= 34) blocs protéiques, nous avons décidé de réduire leur nombre pour posséder différentes séries ayant des nombres de blocs protéiques variables. Ces BPs seront utilisés pour la prédiction. Pour définir les blocs à éliminer,

nous avons décidé d'éliminer les blocs ayant de trop proches voisins. Pour définir le fait que 2 BPs sont proches, deux critères ont été utilisés. Tout d'abord, le fait qu'ils soient proches structuralement, ensuite que leurs transitions soient à peu près identiques. Pour cela, il suffit de calculer le *RMSda* des B PBs pris deux à deux, et de sélectionner les couples dont le *RMSda* est inférieur à une valeur définie. Ensuite, il suffit de faire une différence entre les fréquences de transition des deux PBs et si celle-ci est faible, cela veut dire qu'en effet les deux blocs sont pratiquement équivalents. Le bloc protéique le moins peuplé est alors éliminé.

Ayant réduit le nombre de blocs, il suffit de recommencer l'apprentissage avec le nouvel ensemble de blocs. Le processus s'arrête quand plus aucun bloc ne peut être éliminé avec les critères définis.

3.3 L'alphabet structural

Avec cette approche automatisée, le nombre de blocs passe de $B = 34$ blocs à 22, puis 19, 16, 14, 12, 11 et enfin 10. Le *RMSda* moyen est au départ de 25.4^o. Il augmente à 28.5^o pour 22 blocs puis 29.0^o pour 19 blocs et passe à 30.0^o pour 16 blocs. Il se stabilise ensuite entre 30^o et 32^o. La première réduction est forte, plus d'une dizaine de blocs sont éliminés. Avec la diminution du nombre de blocs protéiques, l'approximation angulaire augmente fortement, puis se stabilise.

Toutefois, pour choisir le nombre optimal de BPs, il faut considérer en parallèle le pourcentage de bonne prédiction lié à chaque série de BPs. Celui-ci est discuté dans le paragraphe 4 (*"Prédiction de la structure locale en blocs protéiques"*).

3.3.1 Description des blocs protéiques

La figure 3.5 montre le résultat de l'étape d'apprentissage avec les angles des 16 blocs. Le tableau 3.1 regroupe leurs valeurs exactes.

La figure 3.6 (visualisé avec le logiciel rasmol et sur la page de garde avec MOLSCRIPT [111]) montre pour chacun des 16 blocs protéiques, numérotés de a à p, des fragments superposés qui donnent ainsi une idée du type de repliement associé à chaque bloc. Les blocs ont été re-ordonnés suivant leurs fréquences de transition ainsi qu'en observant leur répartition dans les structures secondaires définies par la méthode de consensus [31]. Ces informations sont résumées dans le

bloc	ψ_{n-2}	ϕ_{n-1}	ψ_{n-1}	ϕ_n	ψ_n
a	41,14	75,53	13,92	-99,80	131,88
b	108,24	-90,12	119,54	-92,21	-18,06
c	-11,61	-105,66	94,81	-106,09	133,56
d	141,98	-112,79	132,20	-114,79	140,11
e	133,25	-112,37	137,64	-108,13	133,00
f	116,40	-105,53	129,32	-96,68	140,72
g	0,40	-81,83	4,91	-100,59	85,50
h	119,14	-102,58	130,83	-67,91	121,55
i	130,68	-56,92	119,26	77,85	10,42
j	114,32	-121,47	118,14	82,88	-150,05
k	117,16	-95,41	140,40	-59,35	-29,23
l	139,20	-55,96	-32,70	-68,51	-26,09
m	-39,62	-64,73	-39,52	-65,54	-38,88
n	-35,34	-65,03	-38,12	-66,34	-29,51
o	-45,29	-67,44	-27,72	-87,27	5,13
p	-27,09	-86,14	0,30	59,85	21,51

TAB. 3.1 – *Les 8 angles définissant les 16 blocs protéi*

tableau 3.3.

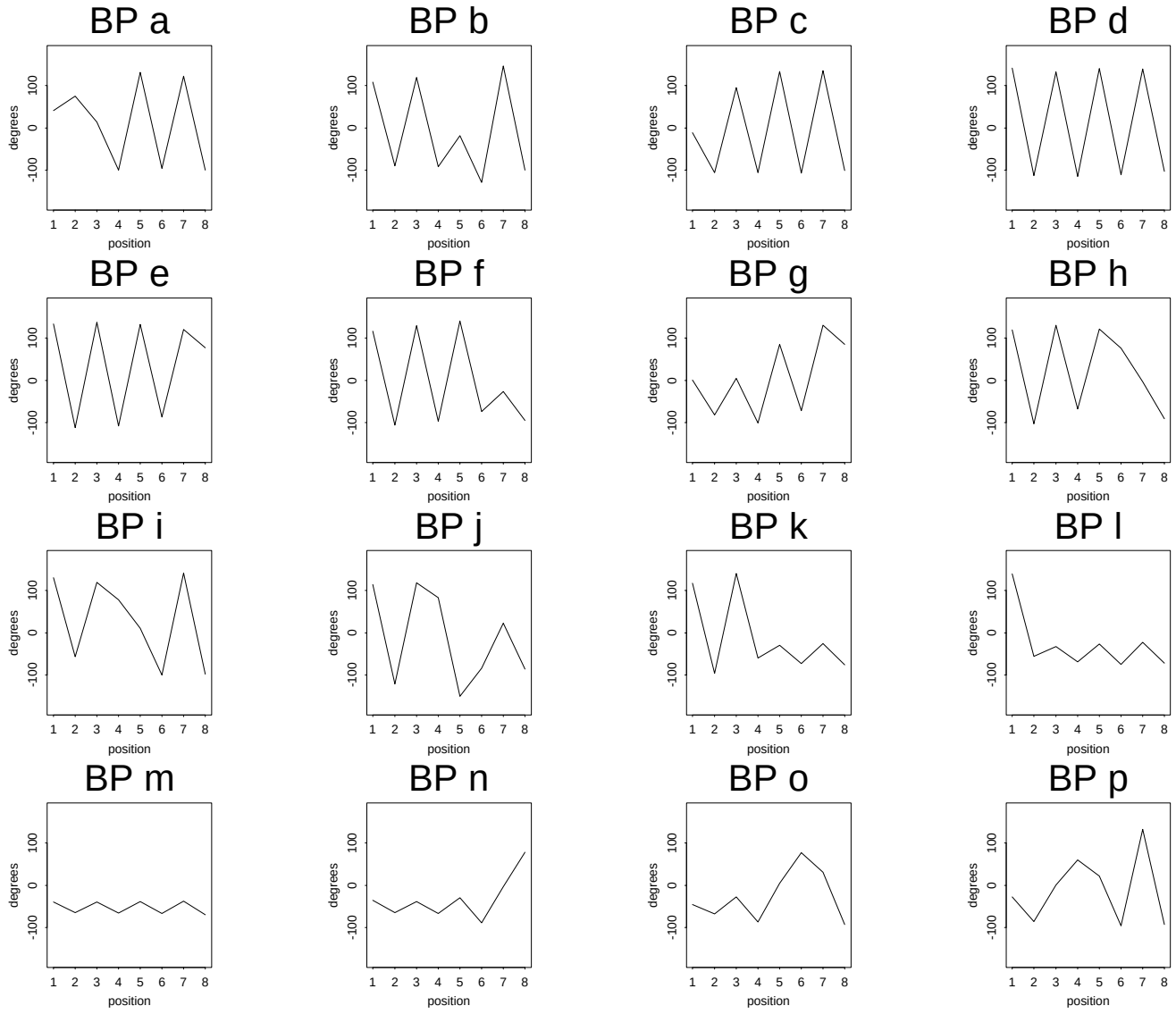


FIG. 3.5 – *Angles des 16 Blocs Protéiques obtenus (en degrés).*

3.3.2 Stabilité structurale des 16 blocs protéiques

3.3.2.1 *RMSda* et *RMSd*

La qualité des blocs a été mesurée en utilisant les critères du *RMSd* et du *RMSda*. Pour le premier, chaque fragment protéique assigné à un bloc donné a été comparé à tous les autres fragments associés à ce même bloc. La moyenne des valeurs observées est récapitulée dans la seconde colonne du tableau 3.3. Ainsi, pour la totalité des blocs protéiques, excepté pour BPj, le *RMSd* est inférieur à 0,74 Å, sur 5 C_α. Il est intéressant de noter que les blocs caractéristiques des structures secondaires, BPm (partie centrale des hélices α) et BPd (pour les feuillets β), ne

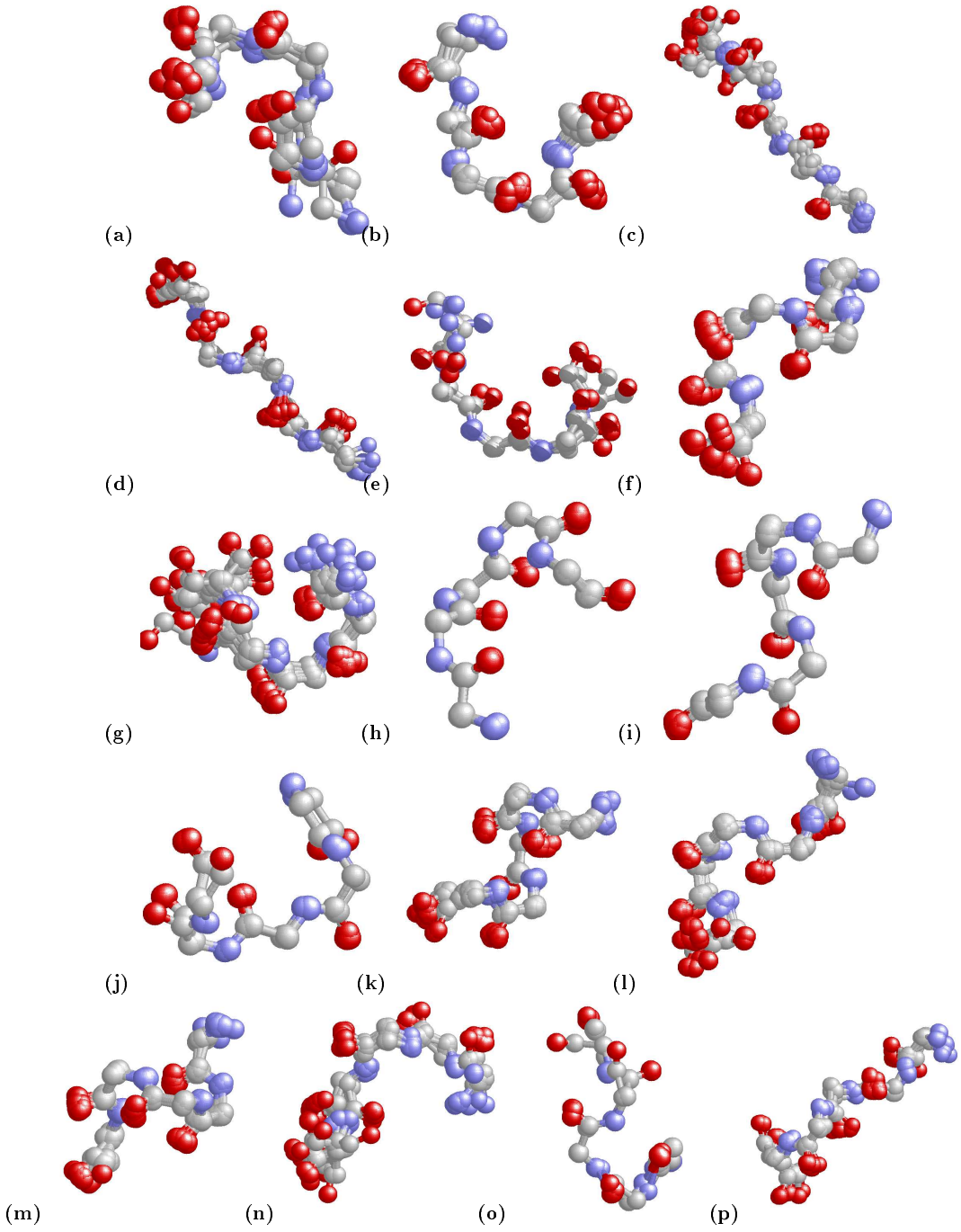


FIG. 3.6 – *Superposition de fragments pour les 16 blocs protéiques.*

sont pas les seuls à être bien approximés, comme le bloc BPp (0.46 Å) et de nombreux autres dont le *RMSd* est inférieur à 0,5 Å.

La moyenne des *RMSda* est de 30.0^o pour l'ensemble des BPs. Seul le BPm a une moyenne individuelle largement inférieure (15^o), les autres sont tous compris entre 25^o et 32^o. Toutefois, il convient de noter la différence qui existe entre la défintion de la moyenne des *RMSda* et de l'approximation angulaire des BPs. La moyenne des *RMSda* est la moyenne de chaque fragment protéique avec le bloc protéique qui lui est associé, or en chaque position (excepté pour les extrémités), chaque angle dièdre est codé non pas par 1 mais par 4 blocs protéiques, les fragments étant chevauchants. Ainsi, en faisant simplement la moyenne des angles des blocs protéiques présents en chaque position, 50 % des angles sont approximés à moins de 21 % de la réalité. Ce point est par ailleurs discuté dans le paragraphe 3.3.2.6, où une nouvelle stratégie est proposée. En outre, la valeur du *RMSda* moyen est particulièrement touchée par quelques valeurs extrêmes qui ont un rôle important dans le calcul d'une moyenne.

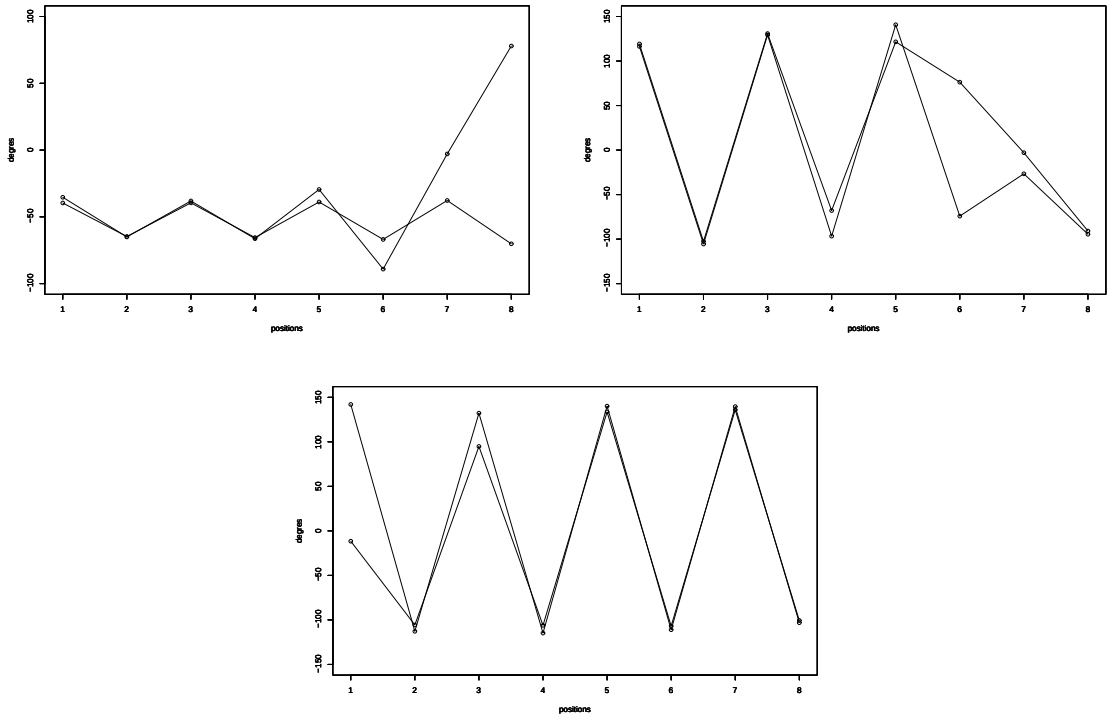


FIG. 3.7 – *Superposition des angles des blocs protéiques ayant le plus faible* RMSda *avec les blocs* m *et* n *(19,2^o),* f *et* h *(19,5^o) et* c *et* d *(19,8^o)*

3.3.2.2 Différence structurale entre les blocs protéiques

Un des points importants est de s'assurer que les blocs sont bien distincts d'un point de vue structural. Les valeurs du *RMSd* calculé entre les blocs vont de 0,21 à 2,07 Å. Les blocs les plus proches sont les blocs protéiques m et n (0,21 Å), f et h (0,23 Å), n et o (0,24 Å), c et d (0,25 Å). Ces blocs sont répartis dans les même types de structures secondaires (cf. tableau 3.3). Le calcul du *RMSda* donne les même associations avec les blocs m et n (19,2^o), f et h (19,5^o) and c et d (19,8^o). Toutefois ces *RMSd* faibles ne traduisent pas les différences réelles. Ainsi le bloc n va vers le BP m qui est une forme très répétitive; ils ont donc une partie commune importante, mais aussi une partie distincte. L'analyse individuelle des angles donne un meilleur éclairage sur ce point.

L'observation des différences angulaires entre ces paires de BPs (cf. figure 3.7) montrent la présence de 5 à 6 angles très proches (moins de 10^o de différence), ce qui induit le *RMSda* et le *RMSd* faible; mais aussi 1 à 3 angles totalement différents (plus de 100^o d'écart). Ce sont ces angles qui donnent leur spécificité au bloc. Il faut noter que le *RMSda* donne ici une information plus précise sur la distinction entre les blocs protéiques, ce qui corrobore les exemples donnés par Unger et collaborateurs (cf. paragraphe 2.3.2.1).

3.3.2.3 Sensibilité de l'assignation par le *RMSda*

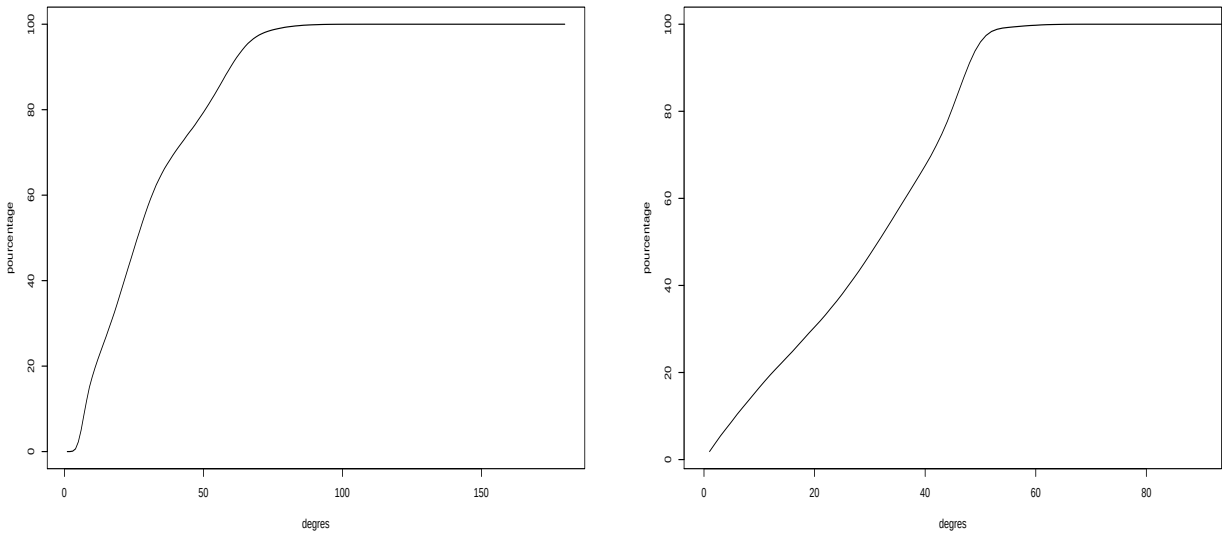


FIG. 3.8 – *(a) Valeur du* RMSda *pour le premier bloc sélectionné. (b) Différence de valeur de* RMSda *entre les second et premier blocs sélectionnés.*

Ayant vu que les blocs sont bien distincts entre eux et possèdent des spécificités angulaires importantes, un second point à observer est la qualité de l'assignation d'un fragment à un bloc. En effet, il faut vérifier que pour une majorité de fragments leur assignation à un bloc protéique est unique. En d'autre terme, il convient d'avoir le second bloc protéique le plus proche, le plus "lointain" possible. Pour évaluer ce critère, il convient donc d'analyser la répartition des valeurs des *RMSda* pour les blocs sélectionnés (blocs réels) et ensuite pour les blocs ayant le second *RMSda* le plus faible. La figure 3.8 récapitule (a) les distributions cumulées des valeurs d'assignations des fragments protéiques à leur bloc et (b) la différence entre le *RMSda* le plus faible (bloc réel) et le second plus faible. Ainsi, la distribution des *RMSda* associés au bloc réel montre que le *RMSda* moyen est faible. Une bonne différence de *RMSda* entre les deux meilleurs *RMSda* est observée. La confusion entre les blocs est donc faible. La différence moyenne et d'une valeur de 40^o. Plus de 80% des blocs en second sont à plus de 20^o de *RMSda*. La distinction entre le bloc "réel" et le second bloc le plus proche est donc bonne, et donc, ceci implique un codage en blocs protéiques de la structure protéique sans ambiguité.

Pour évaluer la sensibilité de l'assignation des fragments par le critère du *RMSda*, une étude de dynamique moléculaire a été menée. Pour cela, des variations du squelette protéique de l'ubiquitine ont été simulées dans les gammes de facteurs de températures classiques (modifications des longueurs inter-atomiques et des angles de valence). Aucun changement d'assignation n'a été observé. L'assignation par le critère du *RMSda* est donc particulièrement stable.

3.3.2.4 Exemple de recodage

La figure 3.9 montre la protéine de liaison à l'ubiquitine (code PDB : 1aak) représentée en 3D par le logiciel MOLSCRIPT [111]. La figure 3.10a la montre recodée en blocs protéiques avec les variations de *RMSd*s en chaque position (cf. 3.10b). Cette protéine est une protéine (α/β). Il est aisé de voir les successions de BPd et BPm. On peut aussi remarquer quatre séries de blocs *cfkl* associées à des boucles amenant à des hélices α, ainsi que quatre séries *dfk*, deux *ehia* et deux *bccd* entre des feuillets β, ainsi que deux séries *opacd* localisées dans les boucles. Le codage global de la protéine est assez correct, seules 4 positions ont un *RMSd* supérieur à 1,0 Å. Les structures répétitives sont bien approximées, mais ne sont pas les seules (cf. paragraphe

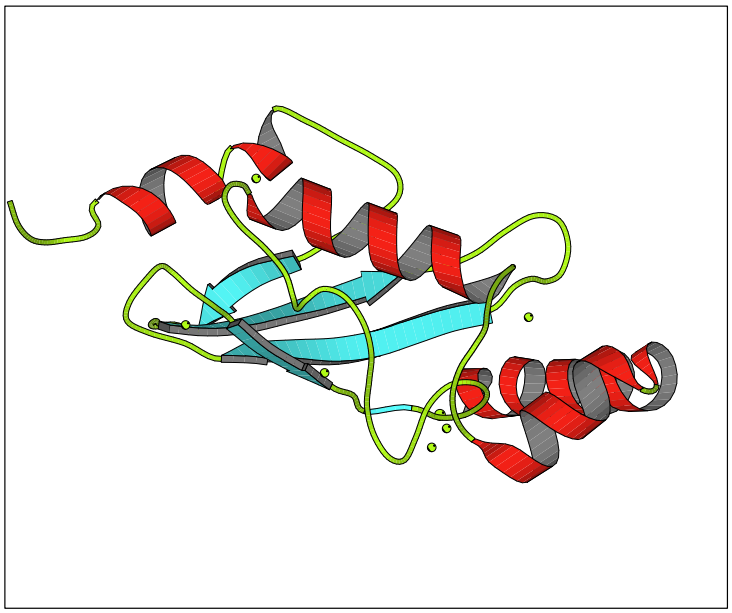


FIG. 3.9 – *Exemple de la protéine* 1aak *représentée avec le logiciel MOLSCRIPT [111]*

motif	nombre d'observations	*RMSd* moyen (Å)
$mm(cc)dd$	30	0,70
$dd(fkl)mm$	414	1,26
$dd(fbdc)dd$	121	1,43
$mm(nopac)dd$	215	0,76
$dd(fkopac)dd$	64	1,05

TAB. 3.2 – *Motifs représentatifs de longueur 2 à 6 entre deux structures secondaires de type BP*m *et/ou BP*d *avec le nombre d'observations pour chaque série et le* RMSd *moyen associé.*

3.3.2.1).

3.3.2.5 Série de blocs: sensibilité structurale

Les autres alphabets structuraux décrits ont souvent des tailles supérieures [162, 53, 19]. Pour analyser la qualité du recodage, des séries de blocs de différentes tailles ont été extraites de la base de données structurales. J'ai regardé spécifiquement des séries de deux structures secondaires répétitives définies par les blocs m et/ou d. Des motifs de tailles 1 à 6 ont ainsi été examinés. Les exemples les plus représentatifs de chaque longueur sont réunis dans le tableau 3.2 avec le *RMSd* moyen associé. Pour expliciter les exemples, $mm(xyz)dd$ est un motif xyz qui relie deux BPm et deux PBd.

Pour des motif courts d'un ou deux blocs, les occurrences sont faibles (moins de 40 observations). Pour des motifs plus longs, le nombre global de combinaisons de blocs protéiques augmente fortement. Par exemple, pour une longueur de 4 (respectivement 5 et 6), un nombre moyen de 20 motifs différents est trouvé (respectivement 22 et 30 motifs). Le nombre et le

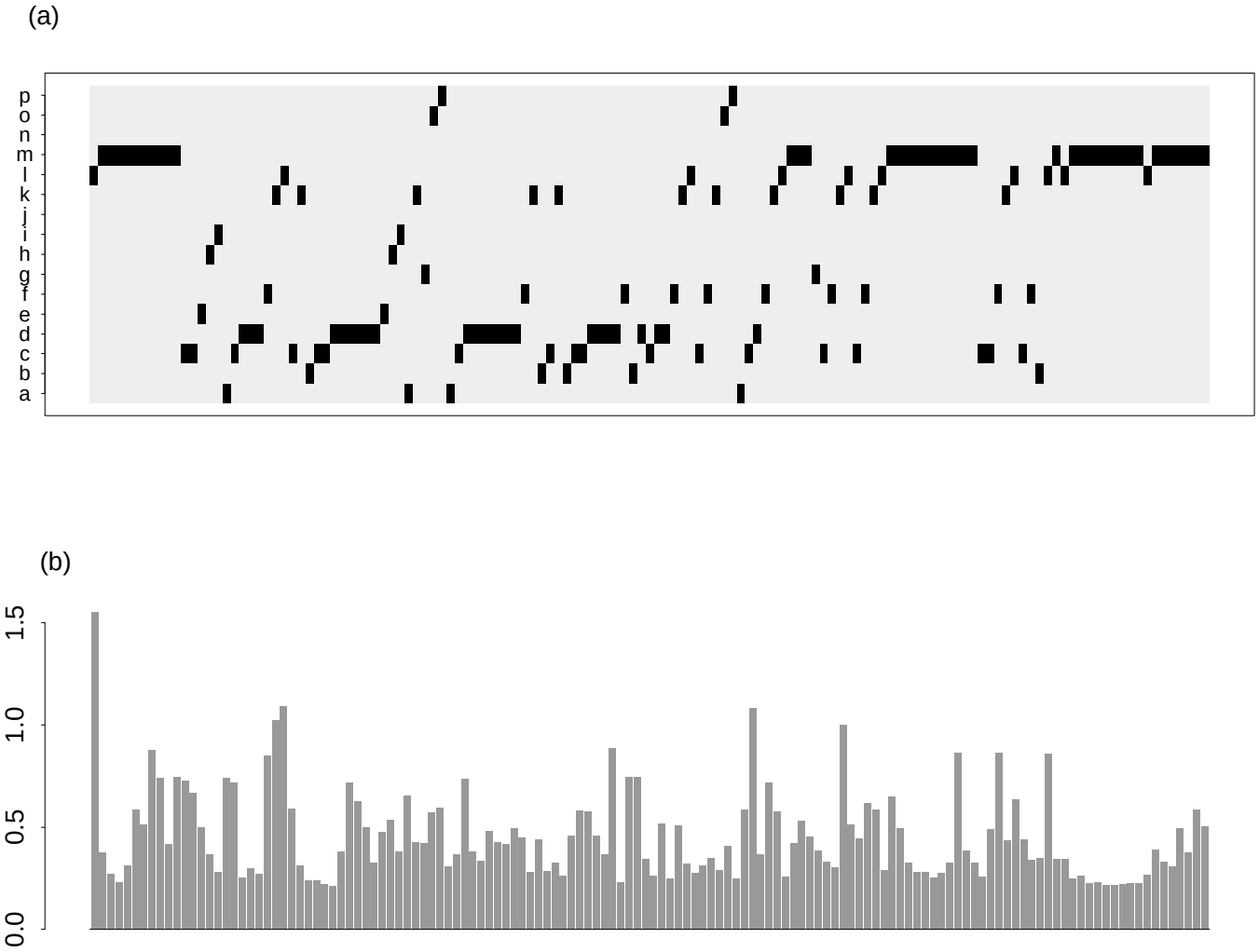


FIG. 3.10 – *(a) Exemple de la protéine de liaison à l'ubiquitine (code PDB : 1aak) représentée en termes de blocs protéiques et (b) les* RMSd *entre chaque fragment de 5 Cα de cette protéine et le bloc protéique auquel il est assigné.*

type de motifs sont importants et dépendent principalement du type de structures secondaires localisées aux extrémités du motif. Pour un motif de longueur 3 ayant en extrémité N-terminale *dd* et finissant par les blocs *mm*, le motif *fkl* représente 98 % des occurrences trouvées; et en reliant *mm* à *mm*, le motif *nop* représente 82 % des occurrences trouvées avec 24 occurrences présentes. Dans les autres cas, et pour toutes les longueurs observées, il n'existe pas de motif représentant plus de 75 % des occurrences.

Les *RMSd* moyens calculés pour ces exemples montrent une forte stabilité structurale avec une moyenne proche de 1,0 Å. L'approximation structurale induite par l'utilisation des blocs protéiques reste correcte pour des motifs de taille importante. Cette notion couplée avec les propriétés des transitions existants entre les blocs, sera reprise dans le paragraphe 5.2 pour des fragments de tailles supérieures à 5 C_α.

3.3.2.6 Reproduction de la structure : amélioration par l'utilisation d'une librairie

Ayant recodé les protéines à l'aide des bloc protéiques, il faut pouvoir ensuite approximer la structure réelle. Pour cela, une première approche simple a été mise au point. Les fragments sont chevauchants donc dans une succession de blocs *tuvw*, tous les PBs *t*, *u*,*v*, et *w* approximent les angles ϕ_u et ψ_t. Ainsi, chaque site, à l'exception des résidus aux extrémités C- et N-terminales, est recodé par 4 blocs protéiques. Une moyenne entre leurs angles, pour le même site, permet un recodage simple. Il faut cependant faire attention aux effets de rotation des angles.

$$\phi_p = \frac{\sum_{i=-1}^{i=2} \phi_{p+i} \ \ modulo \ (360^o)}{4}$$

$$\psi_p = \frac{\sum_{i=-2}^{i=1} \psi_{p+i} \ \ modulo \ (360^o)}{4}$$

Ceci mène à une bonne approximation avec plus de 50% des angles approximés à moins de 21^o et seulement 3 % de mauvaises approximations avec plus de 90^o d'écart avec la réalité.

Pour améliorer ce recalcul des angles à partir du codage en terme de blocs protéiques, j'ai pris en compte le fait que les successions de blocs ne sont pas aléatoires (cf paragraphes 3.3.3 et 5.2). Seul un certain nombre de transitions est vu. Avec une nouvelle base de données de 553 protéines non redondantes [84, 83] et ayant moins de 25% d'identité de séquences (fournie par

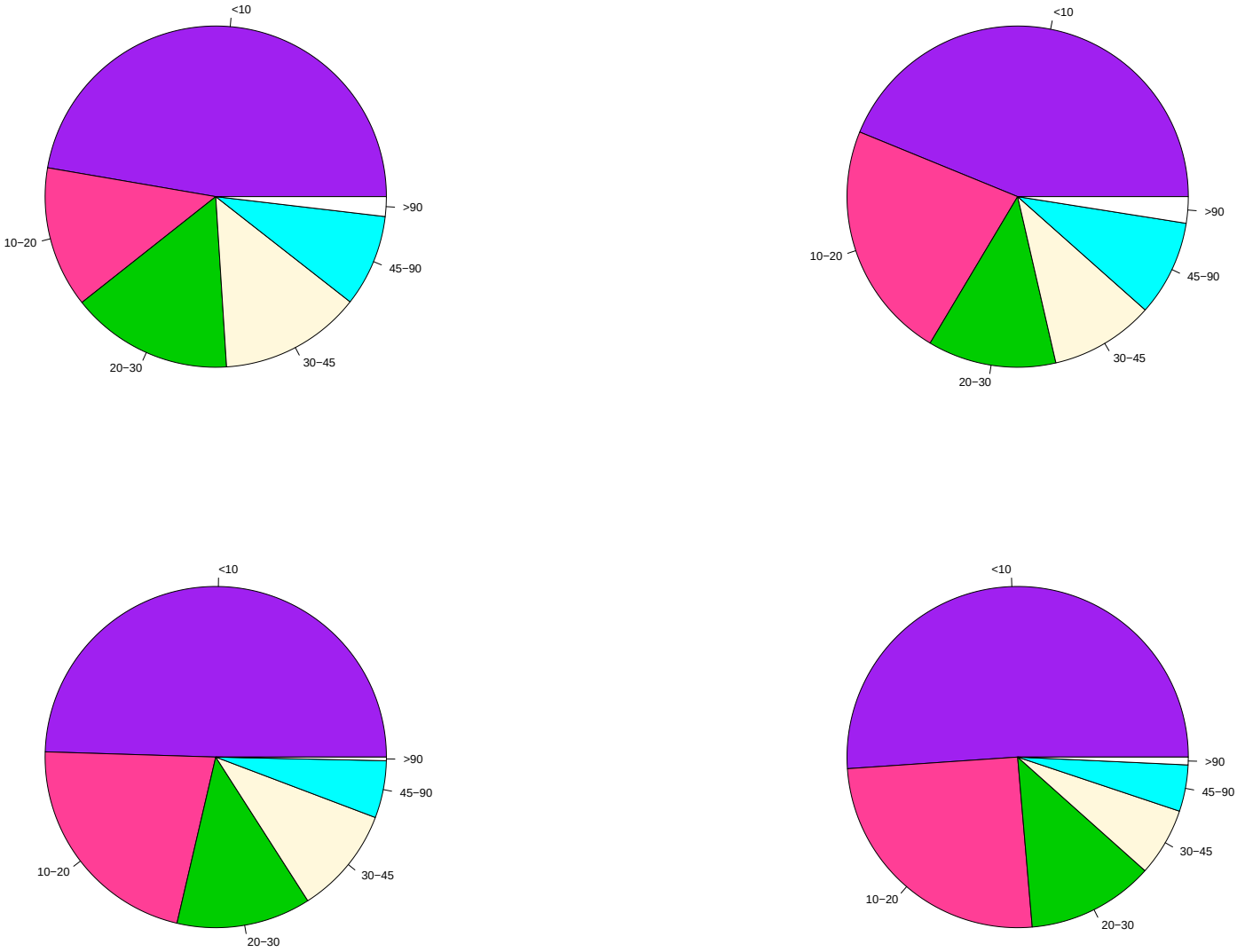


FIG. 3.11 – *Calcul des angles ϕ (respectivement (a) et (c)) et ψ (respectivement (b) et (d))) à partir des blocs protéiques, avec l'usage d'une moyenne des 4 PBs présents (respectivement (a) et (b)), et, avec l'utilisation d'une librairie (respectivement (c) et (d)).*

R. Gautier), j'ai constitué pour chaque succession de 4 blocs indicés par t, u, v, w une librairie avec les angles ϕ_u et ψ_t moyens qui leurs sont associés. Chaque fois que la succession est vue, au lieu de réutiliser la métrique moyenne, les angles moyens ϕ_u et ψ_t associés à cette succession sont pris. Si la succession n'a jamais été vue, l'approximation se fait en réutilisant la métrique précédente. La figure 3.11 récapitule les deux approches. On peut ainsi observer que: (1) d'une manière générale les angles ψ sont les moins bien approximés et ceci avec les deux méthodes et (2) le nombre d'angles mal approximés chute fortement avec l'utilisation d'une librairie. Ainsi, il y a 4 fois moins d'angles à plus de 90^o avec cette seconde méthode et 50% des angles sont approximés à moins de 10^o.

Après avoir observé dans les derniers paragraphes la stabilité structurale de la série de 16 blocs protéiques, nous allons analyser dans les paragraphes suivants les blocs sur un plan structural en observant tout d'abord les transitions qui existent entre les blocs protéiques, puis ensuite la répartition des structures secondaires associées aux blocs et enfin pour mieux comprendre par la suite la prédiction locale, les différentes répartitions des acides aminés dans les blocs protéiques.

3.3.3 Transitions entre les blocs protéiques

Les transitions majoritaires d'un bloc vers d'autres blocs sont en nombre restreint. Elles sont récapitulées dans les cinquième à septième colonnes du tableau 3.3. Les fréquences de transition f_{xy} d'un bloc x vers un bloc y (différent de x) à la position suivante sont calculées par la formule suivante :

$$f_{xy} = \frac{N_{xy}}{(\sum_{z=1}^{z=16} N_{xz}) - N_{xx}}$$

avec N_{xy}, le nombre de blocs x suivi par un bloc y, et N_{xx} suivi par lui-même.

Les trois transitions majoritaires des structures répétitives, en excluant les blocs correspondant aux parties centrales, représentent au minimum 76% de leurs transitions, sauf pour le bloc protéique j. Par exemple, pour le bloc protéique c, 92,2 % des transitions vont vers trois blocs: BP d (62.2 %), BP f (24.4 %) et BP e (5.6 %). De la même manière, moins de la moitié des transitions possibles apparaissent avec une fréquence de plus de 1,0 %. En règle générale, le nombre de transition par blocs dépassant 5 % est de 3, avec un maximum allant jusqu'à 5 pour

le bloc protéique j.

3.3.3.1 Relations avec les structures secondaires

Les blocs protéiques peuvent être caractérisés par leur composition en structures secondaires. Comme pour les autres alphabets structuraux [23], les blocs ne correspondent pas aux catégories classiques des structures secondaires (cf. tableau 3.3). Le bloc protéique m est une partie centrale d'hélice α ($\psi = -47^o$ et $\phi = -57^o$). Le bloc d est un bloc "idéal" de feuillet β ($\psi = 135^o$ et $\phi = -139^o$).

En tenant compte de leurs compositions en structures secondaires et de leurs transitions, les blocs notés de a à c et d à f sont groupés autour du bloc d comme appartenant à des structures β (cf . tableau 3.3). Les blocs k, l et n, o et p sont des structures locales associées à des extrémités N- et C-terminale d'hélices α et avec des taux élevés d'assignation dans des structures α. Le dernier groupe est composé des blocs allant de g à j, et représente le groupe "boucles" (fréquence des boucles supérieur à 80 % pour le C_α central).

La figure 3.12 permet de comparer visuellement l'assignation des blocs par rapport à une assignation par les structures secondaires classiques. Le patchwork est parfois difficilement lisible dans les parties boucles, mais montre bien la finesse de leur attribution. Dans l'hélice α située en bas à gauche, l'utilisation des blocs protéiques permet de bien marquer deux résidus (en bleu) qui sont distincts d'une structure régulière d'hélice. Pour les feuillets, l'intérêt est encore plus grand au vue des différence visible entre le long feuillet β central et le plus court situé à sa gauche qui est plus tordu.

Un autre paramètre a été analysé, le nombre moyen de répétition sur soi-même noté nmr (cf. la quatrième colonne du tableau 3.3). Cette valeur est quantifiée par :

$$nmr = \frac{1}{1 - \frac{N_{xx}}{\sum_{z=1}^{z=16} N_{xz}}}$$

avec $N_{xx}/\sum_{z=1}^{z=16} N_{xz}$ la fréquence de transition du bloc sur lui même, N_{xx} étant le nombre de fois où le BPx etait suivi par lui-même et N_{xz} où il est suivi par BPz. Cette notion permet de bien caractériser les structures répétitives classiques :

(i) BPm possède un taux de transition sur lui même de p_{mm}=85,2%, soit un nmr de 6,74

				Transitions			
BP	f. occ.	RMSdm	nmr	1^{er}	2^{eme}	3^{eme}	α
a	3,93	0,52	1,01	54,8(c)	16,5(f)	8,0(b)	0,1
b	4,58	0,51	1,00	44,4(d)	17,9(c)	13,7(f)	0,2
c	8,63	0,51	1,28	62,2(d)	24,4(f)	5,6(e)	0,1
d	18,84	0,48	2,74	51,9(f)	25,6(c)	19,2(e)	0,0
e	2,31	0,54	1,11	80,4(h)	9,1(d)		0,0
f	6,72	0,50	1,00	60,7(k)	36,3(b)		0,0
g	1,28	0,74	1,05	37,5(h)	28,0(c)	19,1(o)	6,9
h	2,35	0,62	1,04	62,4(i)	18,1(j)	10,2(k)	0,0
i	1,62	0,56	1,01	87,7(a)			0,0
j	0,96	1,03	1,01	17,0(a)	16,6(b)	16,1(l)	3,7
k	5,46	0,59	1,00	76,2(l)	13,6(b)		35,1
l	5,35	0,63	1,01	68,5(m)	9,2(p)	7,0(c)	44,4
m	30,04	0,43	6,74	33,8(n)	18,5(p)	9,7(b)	86,7
n	1,93	0,61	1,03	90,9(o)			68,4
o	2,60	0,60	1,02	74,7(p)	8,3(m)		43,1
p	3,41	0,46	1,00	58,1(a)	22,7(c)	11,1(m)	11,2

TAB. 3.3 – *Pour chaque bloc protéique (noté de BP*a *à BP*p*), sont donnés sa fré (RMSdm), le nombre moyen de répétition sur lui même (*nmr*), ses 3 transitions feuillet β et boucles) pour le résidu central (S2) et une caractérisation grossière*

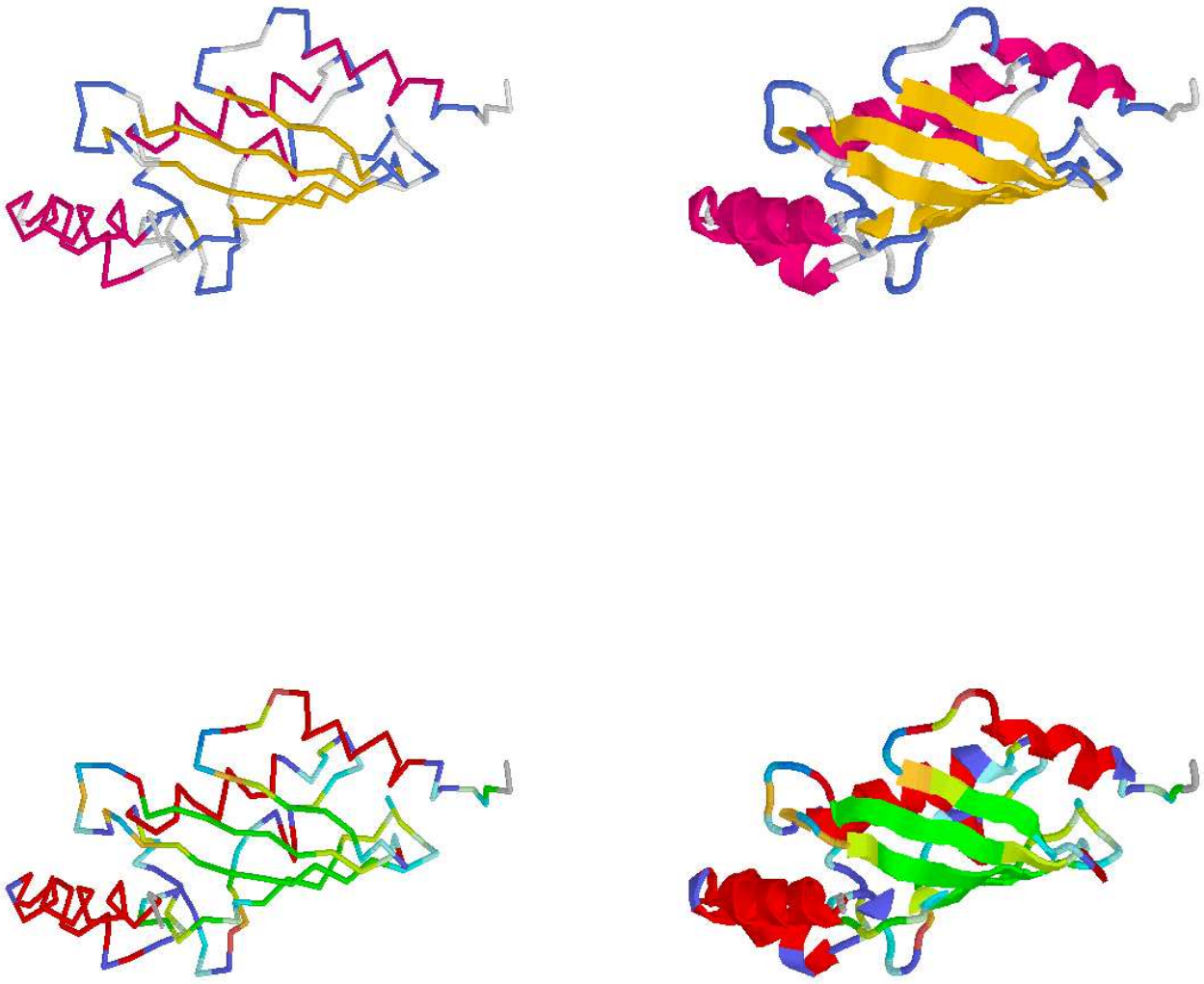


FIG. 3.12 – *Colorisation de la protéine 1aak avec le logiciel rasmol. En haut est représentée la protéine 1aak coloriée avec les structures secondaires, en bas, les même représentations, avec les blocs protéiques.*

blocs ce qui correspond bien à une hélice régulière. De la même manière, 78,1 % des fragments de 5 C_α de la base de données dont le troisième C_α est considéré en hélice α appartiennent à ce bloc et 86,7 % des fragments de ce bloc ont un troisième C_α en hélice α.

(ii) BPd (avec p_{dd} = 63,5 %) caractérise un feuillet β avec une longueur moyenne de 2,74 blocs.

(iii) BPc et BPe ont un nmr supérieur à 1,1. Ils correspondent à des entrées ou des sorties de feuillets, structures de feuillets β déformées ou à des extrêmités C- ou N-terminales de régions étendues du squelette protéique.

La dernière colonne du tableau 3.3 fait une analogie entre les blocs protéiques et les structures secondaires, toutefois cette nomenclature est rapide. Les structures secondaires n'ont que 3 états distincts: hélices α, feuillets β et boucles (ou non-α / non-β), l'équivalence n'est donc que partielle avec un alphabet structural à 16 états. Par exemple, le bloc protéique b mis dans une catégorie "extrémité N-terminale de feuillet β", va vers le bloc d principalement, bloc β par excellence, mais aussi directement vers une "extrémité C-terminale de feuillet β", le bloc protéique f avec un taux de 13,7 %. De la même manière, le bloc m prototype de l'hélice α va vers le bloc protéique b, une "extrémité N-terminale de feuillet β" avec une fréquence de 9,2 %. La flexibilité de l'alphabet est supérieur au simple label donné dans le tableau.

3.3.4 Les acides aminés présents dans blocs protéiques

Après la phase d'apprentissage, l'ensemble des protéines est codé à l'aide des blocs protéiques (pour chacune des séries de B blocs obtenus), en utilisant toujours comme critère d'attribution le $RMSda$ minimal. Chaque bloc est donc associé avec un ensemble de séquences protéiques. Les matrices d'occurrence des différents acides aminés pour chaque position ont donc été calculées. Ainsi, chaque bloc protéique d'une longueur de M=5 résidus est représenté par une matrice de taille M x 20, pour chaque type d'acides aminés. Pour améliorer la spécificité des matrices, le calcul n'a pas été fait sur M résidus, mais pour une fenêtre plus étendue entre -w et +w autour du résidu central. Nous avons testé différentes longueurs w (avec w>2). De manière plus formelle, le nombre d'occurrences n_{ij}^k pour un type d'acide aminé donné (indexé par i =

1,2,...,20) localisé à une position j (j variant dans l'intervalle [-w,+w]) de la fenêtre est calculé. Ensuite, nous en déduisons la probabilité d'avoir ce type d'acide aminé en cette position pour le bloc, P(a_i en j / PB_k) à l'aide du rapport n_{ij}^k / N_k où N_k est le nombre de PB_k de la base de donnée d'apprentissage. P(a_i en j / PB_k) est la probabilité conditionnelle pour l'acide aminé a_i en position j pour le bloc PB_k. Nous avons défini une fenêtre de longueur 15 ($w = 7$). Ce choix est discuté ultérieurement car il dépend des résultats de la méthode de prédiction.

3.3.5 Analyse des matrices d'occurrences des blocs

Pour analyser les spécificités des répartions des acides aminés suivant le type de bloc impliqué, il est possible :

- de quantifier la spécificité de chaque position dans la fenêtre.
- de déterminer quels acides aminés en quelles positions ont une distribution spécifique dans le bloc considéré.

3.3.5.1 La mesure de divergence asymétrique de Kullback-Leibler (KLd)

Pour aborder le premier point, l'entropie relative ou mesure de divergence asymétrique de Kullback-Leibler (KLd, [112]) est particulièrement utile :

$$K(\mathbf{p},\mathbf{q}) = \sum_i p_i \; ln\left(\frac{p_i}{q_i}\right)$$

Ce terme K quantifie la différence existante entre la distribution des acides aminés dans le bloc $\mathbf{p}$: $\{p_i\}_{i=1,...,20}$ et celle attendue au vue de l'occurrence du bloc dans la base de données si tout était aléatoire. Le terme $K_k(\mathbf{p}_j,\mathbf{q})$ a été calculé en chaque position j pour observer la différence existant entre la distribution observée en acides aminés $\mathbf{p}_j$ et la distribution de référence de la base de donnée $\mathbf{q}$ ajustée pour PB_k.

Cette mesure de divergence, notée KLd, permet de détecter des positions "informatives" en observant les positions j dans l'intervalle [-w;+w].

L'entropie relative $K(\mathbf{p},\mathbf{q})$ est une valeur toujours supérieure ou égale à zéro. Elle suit, multipliée par 2N (N étant le nombre d'observations dans la base de données), une loi classique du χ^2 à 19 degrés de liberté (ou ddl, car défini sur les 20 types d'acides aminés). La valeur seuil

a été choisie pour un risque α de premier ordre de 10^{-5}. Toute valeur supérieure à cette valeur seuil est donc fortement significative.

3.3.5.2 Z-scores

Pour évaluer localement l'influence de chaque type d'acide aminé, l'utilisation de la méthode des Z-scores est fort pertinente. Elle permet de mettre en évidence les sous- et sur-représentations. Ainsi, chaque matrice d'occurrence associée à chaque bloc a été normalisée comme suit :

$$Z_{ij}^{k} = \frac{\left(n_{ij}^{k} - n_{ib}\right)}{\sqrt{n_{ib}}}$$

avec n_{ib} le nombre attendu du ième acide aminé ($n_{ib} = \mathrm{N}_k.\mathrm{f}_i$ où N_k et f_i sont respectivement le nombre de PB_k et la fréquence observée de l'acide aminé i dans la base de données). Les Z-scores positifs (ou négatifs) correspondent, pour le bloc protéique k, à des sur-représentations d'acides aminés (ou sous-représentations). La valeur seuil a été prise égale à 4,4 ce qui représente une probabilité p inférieure à 10^{-5}.

3.3.6 Relation entre blocs protéiques et séquences

Les relations entre blocs protéiques et distributions en acides aminés peuvent être analysées en observant les matrices d'occurrences associées à chaque BP.

3.3.6.1 Quatre Exemples

La figure 3.13 montre la structure tridimensionnelle de fragments de 5 C_α associés à 4 blocs protéiques caractéristiques. Ils sont visualisés à l'aide du logiciel XmMol développé par P. Tufféry [194]. Sont donc représentés les blocs protéiques p, b, d et m, ainsi que leurs matrices d'occurrences normalisées en Z-scores (cf. paragraphe 3.3.5.2) et le KLd associé (cf. paragraphe 3.3.5.1).

Les blocs protéiques m, p et d ont une valeur de *RMSd* moyen assez faible (0,43 Å, 0,46 Å et 0,48 Å respectivement). Le PB b est légèrement plus variable avec un *RMSd* moyen de 0,51 Å. Cette variabilité est due principalement à une plus grande variabilité à ses extrémités.

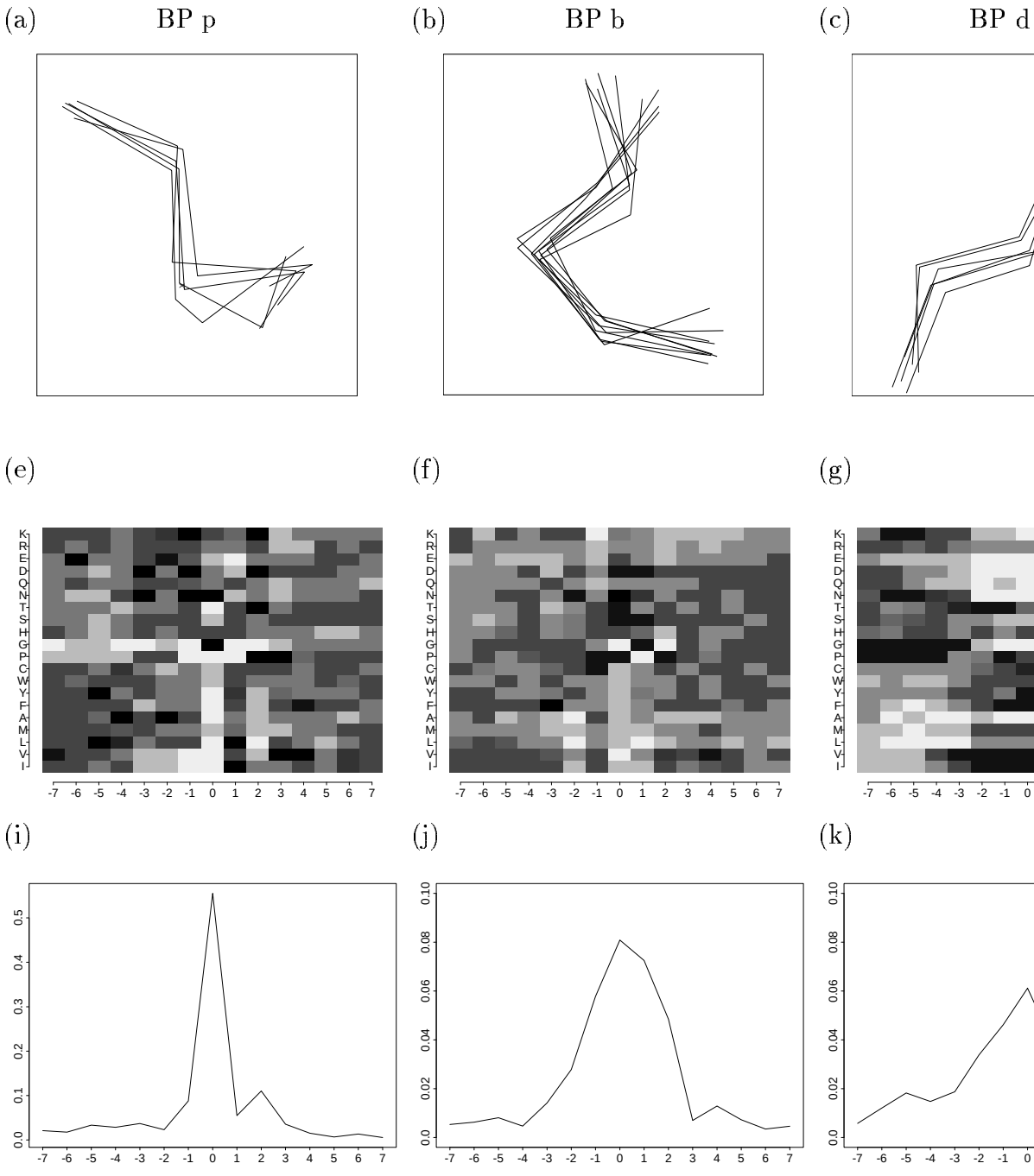


FIG. 3.13 – *Exemple de 4 blocs protéiques. (a,e,i) BP*p*,(b,f,j) BP*b*,(c,g,k) BP*d *(e-h) matrices d'occurrences normalisées en Z-scores, en noir, Z-scores* $> 4,4$*, et (i-l) profils KLd associés à chaque type de blocs.*

L'analyse de la répartition des acides aminés est simplifiée avec l'utilisation des Z-scores. Les matrices ont été seuillées. Ainsi les rectangles noirs indiquent un Z-score > 4,4; inversement pour un rectangle blanc, un Z-score < -4,4. Ce seuil correspond à une erreur de type I avec une probabilité p inférieure à 10^{-5}. Les zones en gris correspondent à un Z-score intermédiaire.

De même, l'analyse du KLd qui exprime la dissimilarité entre la distribution présente dans chaque matrice d'occurrence et la distribution dans la base de données, permet de définir les positions les plus informatives dans un bloc donné. Les 4 blocs protéiques de la figure 3.13 sont représentatifs de l'ensemble des 16 blocs protéiques. Ils ont été ordonnés en fonction de leur valeur de KLd maximale.

Le bloc protéique p est caractérisé par sa position centrale avec une sur-représentation de Glycine et d'Asparagine. Les sous-repésentations sont aussi importantes que les sur-représentations à cette position. Le KLd montre un profil très pointu (KLd = 0,55). Le bloc b possède lui un profil en cloche à fromage cinq fois moins accentué que le précédent avec un KLd maximum de 0,08. En contre-partie de cette diminution de l'importance d'un site, d'autres sites deviennent informatifs dans l'ensemble de l'intervalle [-2:+2] et plus seulement sur la seule position centrale. En parallèle de la diminution des valeurs de Kld, la spécificité en acides aminés devient plus complexe à voir, on peut observer des sur- et sous-représentations alternées de Proline et de Glycine.

Le bloc protéique d correspond à un feuillet β régulier. Le profil du KLd est différent des précédents avec un KLd maximum de 0,06 sur le résidu central et une décroissance symétrique des deux côtés. Une forte différence dans les valeurs de Z-scores entre les Z-scores à l'intérieur et à l'extérieur du bloc structural [-2:+2] est observée. Les sur-représentations concernent principalement certains résidus hydrophobes Isoleucine et Valine et dans une moindre mesure la Phénylalanine, la Tyrosine, le Tryptophane et la Thréonine dans le bloc structural. Les sous-représentations concernent principalement des résidus polaires comme la Lysine, l'Arginine, l'Aspartate, la Glutamine et l'Asparagine; et à l'extérieur du bloc structural la Glycine et la Proline, ce qui est classique [90].

Le bloc protéique m, comme le bloc d, a un profil de KLd particulier. La majorité de la spécificité de la séquence se trouve dans la partie centrale et correspond aux 5 Cα des positions -2 à +2 avec un KLd maximum de 0,05. Nous observons une sous-représentation des résidus aliphatiques, tels que la Leucine, la Méthionine, l'Alanine, et des résidus polaires, tels que

la Glutamine, le Glutamate, l'Arginine, l'Aspartate et la Lysine. La sous-représentation des casseurs d'hélices α que sont la Proline et la Glycine est importante à l'intérieur de la fenêtre structurale. Il en est de même pour l'Histidine et l'Asparagine sur l'ensemble de la fenêtre de 15 résidus [154, 145, 113].

Un autre type de profil existant est un profil bimodal caractéristique de certains blocs, le BPc (maximum en positions -2 et +2), BPe (positions -1 et +1), BPl (positions -2 et 0), BPk (positions -1 et +1) et BPp (positions 0 et +2). Ces positions sont principalement des casseurs de structures répétitives.

3.3.6.2 Utilisation des Z-scores pour déterminer les acides aminés sur- et sous-représentés dans chaque bloc

Le tableau 3.4 mentionne pour chaque bloc protéique, en chaque position, de la fenêtre allant de -4 à +4 autour de la position centrale, les acides aminés sur- et sous- représentés (Z-score >4.4 noté +, ou, <-4.4 noté -). Un grand nombre d'acides aminés par position ont un Z-score important.

Pour se focaliser sur les plus spécifiques des résidus, la valeur de 4,4 comme valeur seuil a été choisie comme pour la figure 3.13. Seule la zone allant de -4 a +4 est représentée. En dehors de cette zone, une seule position présentait encore un intérêt. Les positions informatives ont été retenues sur la base d'une valeur de Kld supérieure à 300 / $2N_k$, avec N_k le nombre de PB_k observé.

Les transitions principales entre les blocs protéiques (cf. tableau 3.3) sont retrouvées dans les compositions en acides aminés. Par exemple, les sur-représentations des Glycines et des Asparagines sont observées pour les blocs protéiques n, o, p et a en positions (+2), (+1), (0) et (-1). De la même manière, les blocs protéiques e et g en position +2, vont vers le BPh (en position +1) and BPi (en position 0). Cependant des différences existent, comme pour le BP d qui va vers le BPf avec une fréquence de 51,9%, sa position (+1) a une sous-représentation différente par rapport au bloc f en position (0) où la Valine et l'Isoleucine sont sous-représentées pour le premier et sur-représentées pour le second. En fait, les proportions des acides aminés ne sont pas toujours conditionnées par les transitions préférentielles entre blocs. Ainsi, la sur-représentation de la Proline en position (+1) du BP b est remplacée par une sous-représentation dans le bloc c alors que la fréquence de transition du BP b au c est de 17,9 %.

BP	Zone		-4	-3	-2	-1	0
a	[-2;-1],[+1,+2]	+		K	DNP	GN	K
		-			GV	AEFILMPSTVWY	GP
b	[-1;+2]	+			N	CP	DST
		-			AL	K	GV
c	(-3),	+		NPS	DGN	HN	IPV
	[-1;+2]	-			AFILMVY	LP	DG
d	[-6;+6]	+	GP	GV	ITV	FIPTV	FIVY
		-	AL	AL	DENQ	ADEGN	ADEGKN
e	(+2)	+		V	V		
		-					
f	[-1;+1]	+			G	FIPVY	DNPST
	,(+3)	-				ADG	AEFGIKLQ
g	(+2)	+					
		-					
h	[0;+1]	+				P	NP
		-					L
i	[-1;0]	+				NP	GN
		-				L	AEILPST
j	(0)	+			G		G
		-					AILPTV
k	[-2;+3]	+		G	FILVY	DNPST	EP
		-			DES	AEFIKLMQRVY	GIV
l	[-1;+2]	+	G	V	DNPST	P	DES
		-			AEFIKLRV	I	ILPV
m	[-4;+6]	+	AEL	ADEQ	AELMQ	AELMQR	AEKLMQ
		-	GP	GPV	GNPST	GNPST	DGPST
n	[0;+2]	+	E	L	AL	KR	AEK
		-		GP			
o	[-1;+1]	+	Y	AL	K	AEK	
		-	P				GPV
p	[-1;+2]	+	A	D		DKN	GN
		-	G		GIPV	AFILPTVY	EGP

TAB. 3.4 – *Les acides aminés les plus importants dans chaque bloc protéique.*
sur-représentations (Z-score > +4.4) et les sous-représentations (Z-score < -4.4

Les sous- et sur-représentations sont concentrées principalement autour de la position centrale, dans une zone recouvrant le bloc structural [-2;+2]. Le tableau 3.4 montre l'importance de ces positions.

On peut noter que les structures répétitives montrent bien des représentations classiques avec des sur-expressions de [AEL] et des sous-représentations de [GPST] pour le bloc protéique BPm, prototype de la partie centrale de l'hélice α, et des sur-expressions de [IV] et des sous-représentations de [ADEGN] pour BP d [90, 145, 113]. De même, la sur-représentation de Glycine est le plus souvent associée à celle d'Asparagine dans les boucles. La flexibilité due à la Glycine permet une torsion importante, et la fonction amine de l'Asparagine permet une liaison directe avec le squelette peptidique assurant ainsi un brusque changement de conformation [47].

3.4 Comparaison avec les autres alphabets structuraux

Une des difficultés majeures de l'évaluation de cet alphabet structural à 16 états est la comparaison avec les autres alphabets existants. En effet, il existe fort peu d'alphabets structuraux accessibles.

Aussi, ce paragraphe présente les différentes études que j'ai pu réalisé. Après une rapide étude avec les structures définies par DSSP, une première comparaison a été effectuée avec les blocs définis par Rooman et collaborateurs [162], que j'ai caractérisé en partant des figures de leur article. La deuxième comparaison a porté sur une base de données recodées par Fetrow et collaborateurs [53], avec leur 6 super-structures secondaires accessibles sur leur site. Et enfin, le dernier alphabet comparé est celui développé par Camproux et collaborateurs [23, 24]; Anne-Claude Camproux m'ayant fourni une centaine de protéines recodées avec son alphabet.

3.4.1 Comparaison avec les attributions de DSSP

Le tableau 3.5 récapitule les sur- et sous-représentations des différentes classes définies par DSSP en fonction des 16 BPs. Les Z-scores sont explicités au paragraphe 3.3.5.2. Les correspondances vues au paragraphe 3.3.3.1 sont retrouvées avec une concentration des hélices α dans les BPs l à n et une sous représentation dans les BPs o à j. Les hélices 3_{10} montrent déjà une plus grande flexibilité d'attribution, elles sont retrouvées avec les BPs g, k, l, et n à p, et surtout jamais avec le BP "hélice centrale" m, ce qui montre encore la sensibilité de l'assignation. Pour

	hélice α	hélice 3_{10}	hélice π	pont isolé	coudes	boucles	feuillet β	conformation étendue β
a	-	-		+	-	+	-	+
b	-	-		-		+	-	+
c	-	-		+	-	+	+	+
d	-	-		+	-	+	+	-
e	-	-		+	-	+	+	
f	-	-		+	-	+	+	-
g	-	+		+		+	-	+
h	-				+	-	-	+
i	-			-	+	-	-	+
j	-				+		-	+
k		+		-	+	-	-	
l	+	+		-	+	-	-	
m	+			-	-	-	-	-
n	+	+		-	+	-	-	-
o	-	+		-	+	-	-	
p	-	+		-	+		-	+

TAB. 3.5 – *Correspondance entre les attributions des structures secondaires par le logiciel DSSP [100] et les 16 PBs, avec (+) Z-scores > 4,4, (-) Z-scores < -4,4 et () Z-scores intermédiaires. La base de données utilisée contient 906 chaînes protéiques possédant moins de 50 % d'identité de séquences.*

les hélices π, aucune n'ayant été observée, aucune corrélation n'est détectable. Ce fait montre aussi l'intérêt d'un alphabet structural qui ne possède pas de structures "rares". Les coudes se trouvent préférentiellement associés aux PBs associés aux hélices α, sauf le PB *m*, sa partie centrale, ce qui semble logique, la plupart des coudes ayant des liaisons hydrogènes proches de ces structures.

Enfin, les boucles et formes β ne se retrouvent pas dans les PBs *k* à *o*, elles sont associées à des BPs différents. Par exemple, les PBs *a* et *b* ne se retrouvent pas dans le feuillet, mais bien dans les boucles et en conformation étendue.

Ce tableau montre bien l'intérêt et la richesse d'analyse possible avec un alphabet structural qui distingue bien les différentes formes classiques entre elles.

3.4.2 Comparaison avec les blocs structuraux de Rooman et collaborateurs

Rooman et collaborateurs ont décrit 4 types de blocs avec 4 longueurs différentes, ici nous n'analyserons que les résultats obtenus pour les blocs de longueurs 5 C_α et 6 C_α. En quelques

SBB	Blocs Protéiques															
	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*	*m*	*n*	*o*	*p*
(a)																
η_5	.	.	.			.	.				+	+	+	.	.	.
ϵ_5			.										.			
ζ_5												.				
λ_5									.	.		.	.			+
(b)																
η_6	.	.	.			.	.	.			+	+	+	.	.	.
ϵ_6		.											.			
ζ_6												.	+			
λ_6												.	.	.		
(c)																
η_6	.	.	.			.					+	+	+	.	.	.
ϵ_6					.											
ζ_6												.	+	.	.	
λ_6													.	+	.	

TAB. 3.6 – *Adéquation entre les blocs structuraux définis par Rooman et collaborateurs[162] et les 16 PBs. Le critère utilisé est celui du* RMSda, *avec comme symboles (+)* RMSda $< 60^o$, *(.)* $60^o <$RMSda $< 75^o$,*et ( )* RMSda $> 75^o$. *(a) Blocs structuraux de longueur 5* C_α, *(b) blocs structuraux de longueur 6* C_α, *sur les 5 premiers* C_α *des blocs, (c) blocs structuraux de longueur 6* C_α, *sur les 5 derniers* C_α.

mots, la famille η_5 (α) est proche des blocs α PBs k à m . La plupart des blocs protéiques associés aux boucles et aux feuillets en sont éloignés. Seule la famille λ_5 (boucles) a une forte association avec le BP p. Les famille ϵ_5 et ζ_5 ne montrent qu'un fort éloignement par rapport aux BPs.

Le bloc structural de longueur 6 C_α a été analysé en deux parties pour les 5 premiers et les 5 derniers C_α. Comme précédemment, peu de lien existe entre un bloc ou ensemble de BPs et l'un des 4 blocs structuraux. La famille η_6 correspond à une hélice α possédant une extrémité N-terminale, alors que la famille ζ_6 est synonyme de la partie centrale de l'hélice. λ_6 correspond bien à des boucles, et est proche du BP n. La famille ϵ_6 (β) n'est pas aussi liée aux blocs β que cela.

De cette simple comparaison, on peut conclure qu'il n'y a pas de correspondance directe entre ces deux types d'alphabet. Une première raison est la différence dans le nombre de blocs utilisés (4 contre 4 fois plus). Une autre cause est sûrement le faible nombre de protéines utilisées dans la première étude. Il semble que ce sont certains repliements qui ont été appris, ainsi la figure 2.13b montre clairement que pour ζ_6, censée être caractéristique des boucles et des feuillets β, le bloc le plus proche est le bloc m soit une hélice α. Cette famille est donc caractéristique non pas de boucles et des feuillets β simples, mais de boucles et des feuillets β allant vers une

	Blocs Protéiques															
blocs	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*	*m*	*n*	*o*	*p*
η		5,7						**11,4**	1,6	2,2	**11,9**	9,2	23,1	15,4	11,7	2,7
α												**10,6**	**86,1**			
τ	7,1	6,4	**10,2**	**16,0**		2,2	5,0		12,7			5,2	7,3	1,2	**8,2**	**16,4**
ζ	2,0		1,5	1,3	9,4	**34,9**	1,9	5,1		2,1	**27,2**	2,7	9,0			1,1
ι	7,1	**10,6**	**15,8**	**40,4**	6,5	3,7	2,0	1,9	1,5	1,1	1,0	1,0	1,2			4,9
β	7,2	7,1	**18,3**	**53,2**	2,4	6,1		1,6								1,9

TAB. 3.7 – *Correspondance des 6 blocs définis par Fetrow et collaborateurs [53] avec les 16 BPs (seules les fréquences supérieures à 1,0 % sont notées, celles en gras sont supérieures à 10,0 %).*

hélice α. Ayant fait des apprentissages avec un nombre aussi limité de blocs, le résultat est surprenant. Pour avoir ce type de blocs caractéristiques, il faudrait au moins deux fois plus de blocs. La seule possibilité est que les protéines utilisées aient été particuliérement riches en certains repliements particuliers.

3.4.3 Comparaison avec les super-structures secondaires de Fetrow et collaborateurs

Le tableau 3.7 montre la correspondance entre les 6 blocs structuraux décrits par Fetrow et collaborateurs [53] et les 16 blocs protéiques. La comparaison est faite en observant la correspondance entre les protéines codées avec les 6 blocs structuraux (partie centrale, 5 C_α de 7 utilisés) et ces protéines codées en 16 blocs protéiques.

Le bloc structural α est concentré sur les blocs protéiques *l* et *m* (96,7 %) et correspond donc à une hélice α et une extrémité N-terminale d'hélice α. Les autres blocs structuraux sont moins bien caractérisés. Le bloc structural η est plus flou et représente surtout un bloc non-β avec une certaine tendance hélice α. A l'inverse, le bloc β est surtout un bloc non-α complet. Le bloc ι possède une forte similitude avec ce dernier. Toutefois, il est moins régulier car il a une plus forte propention de BPs "boucles" (PB *g* à *j*) et surtout il a l'entrée β N-terminale PB *b*. Le bloc structural ζ est encore un bloc lié au feuillet β, moins caractéristique. Seul le bloc structural τ représente bien les "boucles".

Tout comme pour le précédent exemple, aucune correspondance directe n'est envisageable, du fait de la différence du nombre de blocs et des tailles utilisées. Cependant, il semble que cet alphabet soit assez pauvre, mettant le poids principalement sur les structures répétitives et limitant ainsi le rôle des boucles. Une explication possible est l'utilisation de différents types

PBs \ SBBs	α_1	α_2	α'	α'_+	α'_-	$\gamma_{\alpha\beta}$	$\gamma_{\beta\alpha}$	γ_1	γ_2	γ_β	β_2	β_1
a		0,2		2,1	0,2	11,4	0,4	3,5	0,6	3,9	0,6	**77,0**
b		6,2	2,3	**47,6**	0,4	8,8	**26,1**	4,9		0,8	1,5	1,5
c		0	0,5	2,6	0,2	1,7	2,2	0,7	**21,9**	9,4	**29,1**	**31,7**
d		0,1		1,3	0,1	0,3	1,1	0,9	16,4	9,3	**55,8**	14,7
e		0		7,7		0,4	3,5	2,1	18,3	15,5	**45,1**	7,4
f		0	0,1	4,3		0,6	5,6	1,2	**36,8**	8,4	**27,6**	15,3
g		10,7	19,7	4,1	1,6	8,2	3,3	1,6	2,5	1,6	5,7	**41,0**
h		0,7		6,8	3,9	0,7	2,5	11,1	**22,9**	**23,3**	17,6	10,4
i			0,9		18,9	0,5	0,9	**76,5**		0,9		1,4
j					2,8			10,4	2,8	**54,7**	0,9	**28,3**
k		0,2		**31,1**	0,3	2,3	**55,0**	9,8	0,2	0,2	0,5	0,6
l	10,7	**64,1**	0,8	1,8	0,9	9,2		5,4		0		7,2
m	**65,1**	18,4	6,6	1,1	0,1	4,4	0,4	0,9		0	0,1	2,9
n	**35,3**	**27,5**	14,9	3	1,5	7,1	2,2	4,1	0,4	0,4		3,7
o	**24,8**	**31,9**	**20,5**	1,4	2,8	5,7	0,9	6	0,3			5,7
p		15,9	7,6	2,2	**55,4**	4	0,4	4,3	0,7	0,2	0,7	8,5

TAB. 3.8 – *Correspondance entre les 12 blocs protéiques (SBBs) définis par des chaînes de Markov cachées par Camproux et collaborateurs [23] (centré sur le troisième C_α) et les 16 blocs protéiques (centrés sur le troisième C_α). Les fréquences données sont relatives à chacun des 16 blocs.*

d'informations.

3.4.4 Comparaison avec les blocs structuraux de Camproux et collaborateurs

Les deux précédents exemples montrent bien l'intérêt d'un alphabet plus complet avec un nombre d'états supérieurs à une dizaine. La comparaison avec l'alphabet à 12 états mis au point par Anne-Claude Camproux qui a utilisée la méthode des Chaînes de Markov Cachées est donc particulièrement intéressant. L'analyse est en deux parties. La première étant simplement la recherche de l'équivalence entre le bloc protéique et le bloc structural. La seconde tient compte du fait que les blocs structuraux (SSBs) font 4 C_α, et donc un bloc protéique est représenté par une série de deux blocs structuraux.

Les répartitions des blocs sont nettement plus déterministes qu'auparavant. Par exemple, le bloc structural α_1 ne correspond qu'aux blocs associés aux hélices (PBs *l* à *o*). Plus de 80% du BP *m* se retrouve en α_1 et en α_2, et inversement, le BP *d* ne leur est jamais associé. Toutefois, pour les blocs correspondant de part et d'autres aux boucles, les situations sont nettement plus

diverses. Le bloc structural γ_1 est principalement lié aux PBs h à k, alors que le bloc structural $\gamma_{\alpha\beta}$ est réparti dans un nombre nettement plus grand de blocs protéiques. De même, les blocs associés aux feuillets sont particulièrement contrastés. Ainsi, le PB a est lié principalement au bloc structural β_1, mais le bloc structural β_1 est lui présent dans de nombreux PBs distincts (PBs a, c, d, $f$et j).

Les blocs protéiques associés aux hélices α sont les blocs les plus directement reliables aux blocs structuraux avec un nombre limité de successions préférentielles et surtout quelques unes fort importantes. Le PB m est totalement compris dans les blocs structuraux α_1 et α_2.

Cette étude montre qu'il y a une convergence certaine entre ces deux approches malgrè des informations et la méthode différentes, les alphabets obtenus possèdent une grande spécificité. Leur utilisation conjointe serait d'un grand intérêt.

3.5 Conclusion

Pour clore cette première partie sur l'alphabet structural, plusieurs points ont été vus. Tout d'abord sur la méthode utilisée qui tient compte d'un aspect statistique simple avec une prise en compte de l'aspect séquentiel des protéines.

Ayant testé un nombre important de série de blocs, la série de 16 blocs protéiques a été conservée. Elle montre une approximation particulièrement correcte de la structure tridimensionnelle des protéines. De plus, les blocs sont bien distincts entre eux et les fragments protéiques qui leurs sont associés ne sont proches réellement que d'un seul bloc. Ils permettent un recodage particulièrement efficace des structures 3D, avec plus de 50 % des angles approximés à moins de 10^o. L'utilisation du *RMSda* comme critère de sélection a aussi démontré son efficacité. La comparaison avec les autres alphabets met en évidence la pertinence d'un choix de plus d'une dizaine de blocs.

En outre, la grande variété des BPs permet de mieux analyser la structure protéique que la structure secondaire et d'observer des distributions informatives en acides aminés. La quantification de cette information montre la spécificité de nombreux blocs, nous allons donc utiliser cette information dans le chapitre suivant avec une méthode de prédiction de la structure locale à partir de la séquence.

Chapitre 4

Prédiction de la structure locale en blocs protéiques

4.1 Objectif

L' alphabet structural défini permet d'analyser de manière fine la structure tridimensionnelle des protéines (cf. paragraphe 3, *"Apprentissage de la structure locale du squelette protéique"*). La répartition en acides aminés dans chaque bloc montre une forte spécificité (cf. paragraphe 3.3.4); cette information paraissant pertinente, nous allons l'utiliser directement dans une méthode de prédiction de la structure locale à partir de la séquence.

La prédiction a été effectuée avec une méthode bayésienne proche de celle utilisée pour la prédiction des structures secondaires [193], de l'accessibilité au solvant [192] et de modèles plus théoriques d'analyse des modes biologiques [117].

Dans un second temps pour améliorer le taux de prédiction nous avons pris en compte le fait que différentes séquences peuvent être liées à un même type de repliement (*1 bloc protéique* $\rightarrow$ *n séquences*).

Enfin deux stratégies ont été développées pour proposer en fonction d'un taux de prédiction moyen un certain nombre de blocs potentiels. En effet, un type de séquence n'est pas toujours associé au même type de repliement (*1 séquence* $\rightarrow$ *n blocs protéiques*). Ces derniers travaux se basent principalement sur un indice entropique tenant compte de la qualité de la prédiction locale.

Il convient de noter que pour définir la méthode de prédiction, nous avons décidé d'utiliser une méthode statistique simple qui permet de comprendre quels acides aminés sont importants pour chaque type de structures. Il est certain que les réseau neuronaux donnent des résultats

meilleurs, mais en contrepartie, il est peu aisé de comprendre comment ils ont "appris". Aussi, une approche de type bayésienne est particulièrement apropriée.

4.2 Prédiction bayésienne simple

4.2.1 Méthodes

Pour chaque site s d'une protéine, qui comprend aussi bien la position centrale que l'ensemble de la fenêtre de la séquence [-w;+w] autour de cette position centrale, nous avons calculé pour une séquence d'acides aminés X_S, la probabilité d'observer cette séquence dans un bloc donné PB_k, notée $P(PB_k/X_S)$.

De cette probabilité conditionnelle préalablement définie, il est possible de calculer la probabilité d'avoir ce bloc connaissant la séquence, en utilisant le théorème de Bayes. Il accomplit, en effet, l'inversion de la séquence X_S et de la structure PB_k:

$$P\left(PB_k/X_S\right) = \frac{P\left(X_S/PB_k\right).P(PB_k)}{P\left(X_S\right)}$$

avec $P(PB_k)$, la probabilité d'observer PB_k dans la base de données et $P(X_S)$, la probabilité d'observer la séquence d'acides aminés X_S sans aucune information sur la structure. Cette dernière est égale au produit des fréquences des acides aminés dans la base de données. Une approche assez similaire a été utilisée par Thompson et Goldstein [193] pour la prédiction des structures secondaires.

Le terme $P(X_S/PB_k)$ est la probabilité conditionnelle d'observer une séquence donnée X_S (a_{-w},...,a_{+w}) pour un bloc PB_k. Il est calculé comme le produit des probabilités pour chaque acide aminé en position j dans la séquence dans le bloc (cf. figure 4.2). Ce qui amène à l'équation:

$$P(X_S/PB_k) = \prod_{j=-w}^{j=+w} P(a_j/PB_k)$$

Pour définir le bloc optimal PB^* pour une série d'acides aminés X_S en un site s d'une protéine, nous utilisons le ratio R_k (ou son logarithme) défini par:

$$R_k = \frac{P\left(PB_k/X_S\right)}{P(PB_k)} = \frac{P\left(X_S/PB_k\right)}{P(X_S)}$$

Du théorème de Bayes, R_k est défini par le ratio $P(X_S/PB_k)/P(\mathrm{X}_S)$ qui est calculé à partir des matrices d'occurrences. Grâce à ce ratio, la probabilité d'observer un bloc donné PB_k sachant la séquence X_S est comparé à la probabilité d'observer PB_k sans avoir d'information sur la séquence. Ainsi, quand $ln(R_k)$ est positif, la connaissance de la séquence X_S est favorisée par les occurrences de PB_k, et inversement quand il est négatif.

La règle pour définir le bloc optimal PB^* pour la séquence X_S revient à sélectionner, parmi les B blocs, le bloc PB pour lequel ce ratio R_k est maximum. Par conséquence, une liste des B blocs protéiques est définie selon leurs valeurs décroissantes de R_k, le bloc optimal étant le premier. Ainsi, nous pouvons calculer le pourcentage de bonne prédiction Q(1) au premier rang et and Q(r) quand le bloc réel est parmi les r premières solutions.

4.2.2 Résultats

4.2.2.1 Choix du nombre de blocs

La prédiction bayésienne a été effectuée pour chaque série de blocs obtenue, allant de $B =$ 34 blocs, à 22, puis 19, 16, 14, 12, 11 et enfin 10 blocs. La taille de la fenêtre de prédiction a été prise égale à 15 résidus, soit 5 de part et d'autre du bloc structural. La figure 4.3 récapitule les résultats obtenus pour la série de 16 blocs avec des tailles de fenêtre allant de 5 à 19 résidus. A partir de 15 résidus, il y a une saturation dans le gain du taux de prédiction, des résultats similaires ont été obtenus pour les autres séries. Comme attendu, plus le nombre B de blocs augmente, plus le taux de prédiction diminue (cf. figure 4.3).

Dans le choix du nombre de blocs conservés, deux séries (11 et 18 PBs) ont été enlevés car ils avaient un taux de prédiction inférieur à des séries ayant plus de blocs protéiques. En observant les différentes séries obtenues (cf. figure 4.4), on s'aperçoit qu'avec peu de bloc ($B = 10$), le taux de prédiction est bon (39 %), mais l'approximation structurale est alors plus faible (*RMSda* moyen de 32^o). Le choix de 16 est le plus approprié car le taux de prédiction est acceptable (34 %), le *RMSda* moyen reste correct (30^o). En outre, le bloc le moins représenté est égal à un pour cent de la base de données. Cette dernière remarque a son importance: pour la série précédente, les blocs les moins observés représentent moins de 0,5 % de la base de données et il est donc difficilement utilisable pour la prédiction (le nombre d'observations étant alors trop faible).

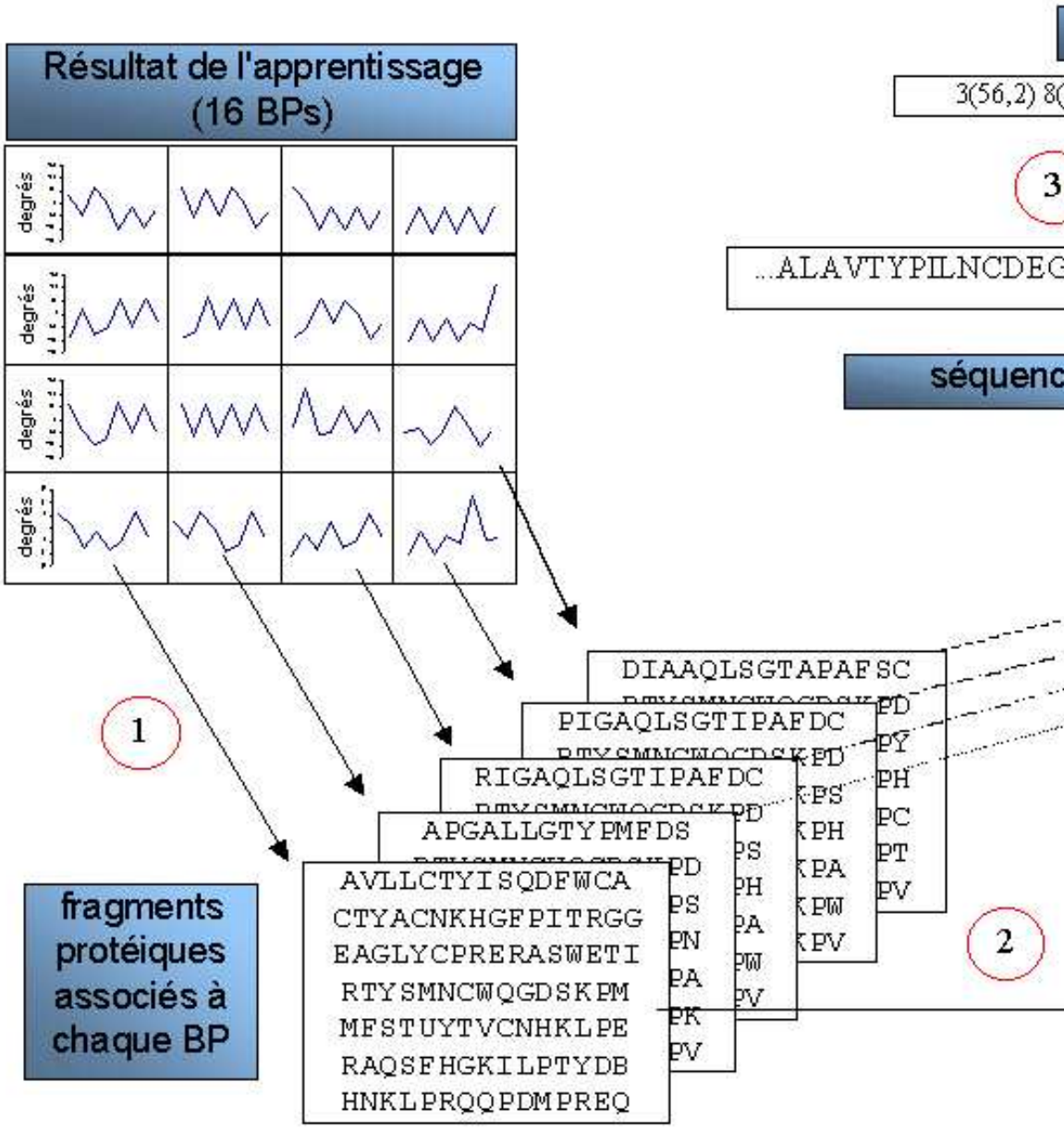


FIG. 4.1 – *Schéma de la prédiction bayésienne: (1) les fragments protéiques sont as*
15 × 20 est normalisées en fonction de la fréquence de chaque type d'acides am
sont classés par ordre décroissant.

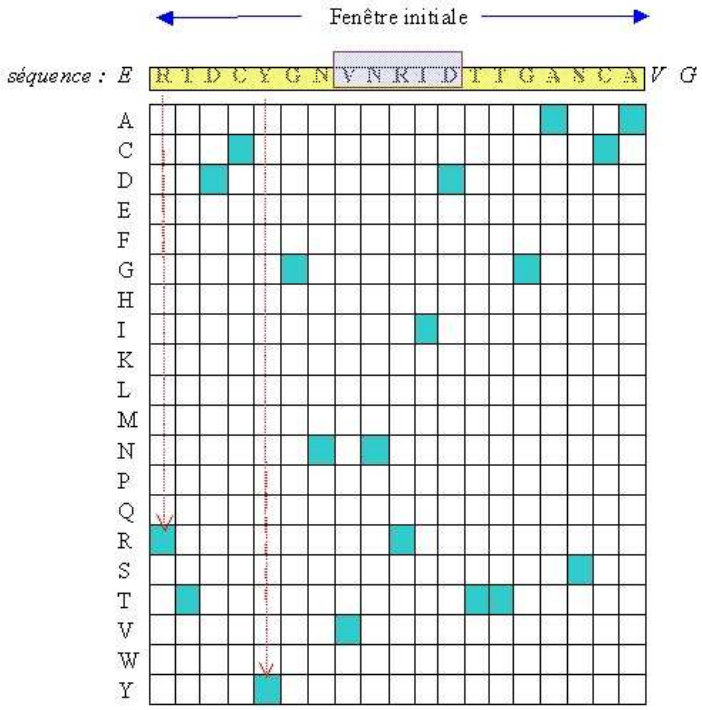


FIG. 4.2 – *Principe du calcul du score. En chaque position de la séquence, le produit des fréquences normalisées est effectué avec chaque matrice d'occurrence normalisée pour chaque BP.*

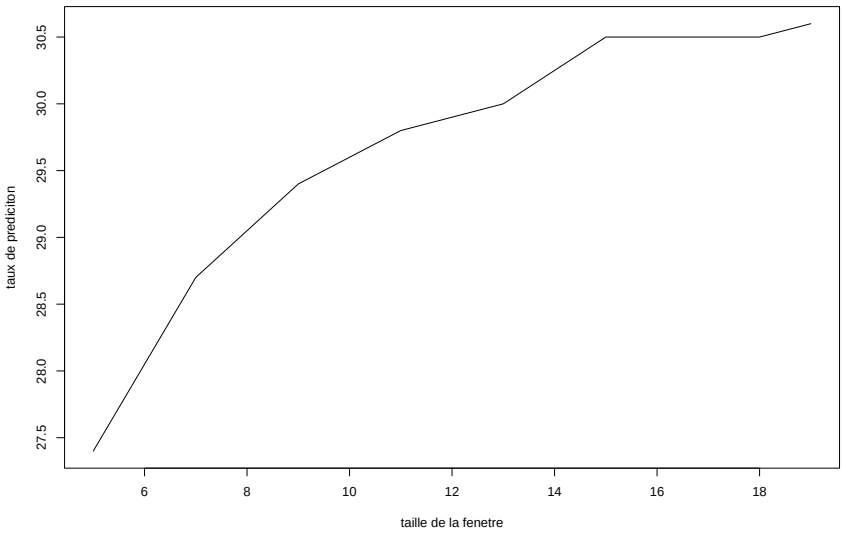


FIG. 4.3 – *Evolution du taux de prédiction en fonction de la longueur de la fenêtre pour la série de 16 blocs protéiques, avec la taille de la fenêtre de prédiction, en abcisses et le pourcentage de prédiction associé, en ordonnés.*

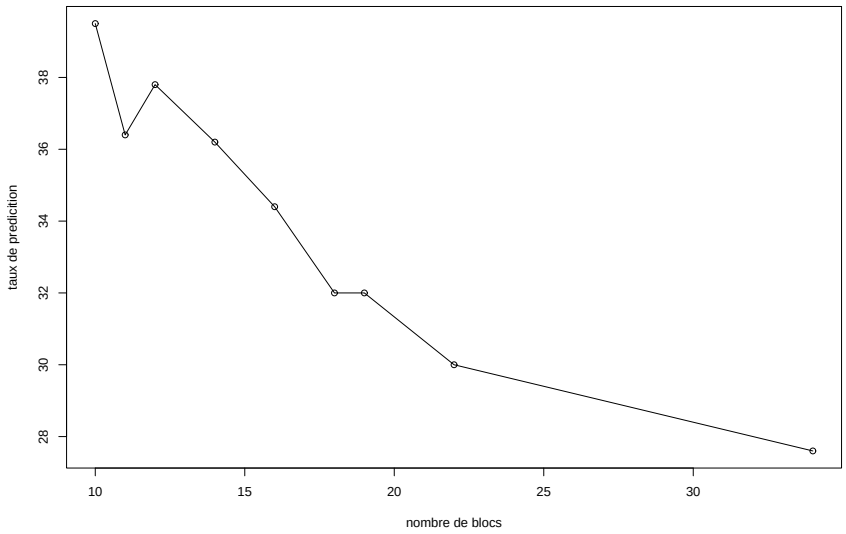


FIG. 4.4 – *Evolution du taux de prédiction en fonction du nombre de blocs, avec le nombre de blocs protéiques dans chaque série, en abcisses, et le pourcentage de prédiction associé, en ordonnés.*

4.2.2.2 La prédiction locale

En utilisant la stratégie bayésienne simple, avec une fenêtre de 5 résidus, correspondant donc au bloc structural, le taux de prédiction est de 30,0 %. Avec une fenêtre totale de 15 résidus, soit 5 de part et d'autre du bloc, le taux de prédiction passe à 34,4%. On peut noter qu'une recherche purement aléatoire donnerait un taux de prédiction de 7 %. L'ensemble des blocs protéiques gagnent en gain de prédiction, par exemple, BPb passe de 11,0 % à 13,5 %, BPe de 33,0 % à 43,2 %,BPi de 32,9 % à 42,2 %, et BPp de 26,9 % à 33,5 %.

Une certaine hétérogénéité est observée dans les taux de prédictions associés à chaque bloc protéique qui va de 13,3 % pour le bloc b à 60,3 % pour le BP a. La figure 4.5 récapitule les différents taux de prédiction par blocs, et, montre aussi le taux de prédiction obtenu quand un certain nombre de solutions sont conservées. En effet, en classant les blocs par score décroissant, le bloc réel se trouve assez réguliérement parmi les blocs les plus probables, ayant donc un score élevé. Ainsi, en conservant les 2 blocs les plus probables, le pourcentage d'avoir le bloc réel passe à 52,1 %, pour 4 solutions conservées, le taux est à 71,4 % et avec 8 solutions conservées sur les 16, soit la moitié, le taux passe à 90,4 %.

Cette distribution montre bien que l'utilisation d'une information séquentielle contient assez d'information pour conditionner la structure locale.

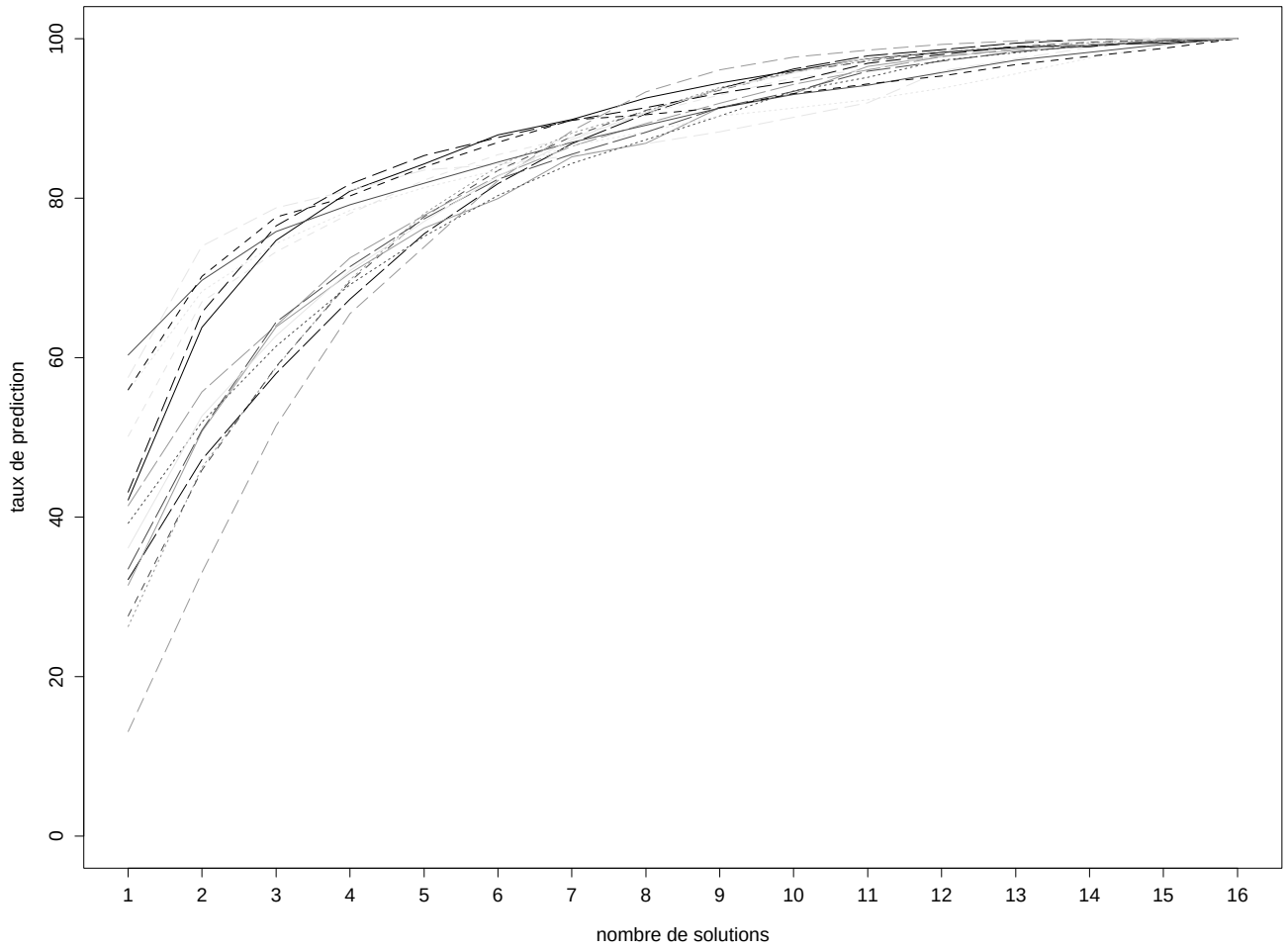


FIG. 4.5 – *Evolution du taux de prédiction pour les 16 blocs protéiques en fonction du nombre de blocs sélectionnés par l'approche bayésienne.*

4.3 Les familles séquentielles

4.3.1 Principe des familles séquentielles

La méthode de prédiction bayésienne implique pour chaque bloc l'utilisation d'une matrice d'occurrence qui lui est propre. Ainsi, si un bloc est composé de deux types de séquences distinctes, l'utilisation d'une matrice commune entraîne un phénomène de moyennage qui fait perdre de l'informativité à ce bloc. Aussi pour améliorer la prédiction, nous avons mis au point la méthode "des familles séquentielles". Elle consite en la génération, pour un bloc protéique x de f matrices distinctes contenant chacune une partie des fragments protéiques du BP x. Pour cela, nous avons procédé à une classification des fragments par une méthode proche des cartes topologiques de Kohonen [109].

Pour un bloc protéique x, f matrices d'occurrence sont créées. Chacune des f matrices est initialisée en mettant les fréquences des acides aminés de la matrice associée au bloc x avec une légère variation pour les individualiser. Toutes les fréquences ont été recalculées pour avoir une fréquence égale à 1 en chaque position.

Ensuite, chaque fragment protéique associé au BP x est alloué à la matrice qui lui ressemble le plus parmi les f matrices. Pour cela, la probabilité conditionnelle $P_l = P(X_S/BP_k^l)$ est calculé pour l allant de 1 à f, avec X_S la séquence en acide aminé correspondant au fragment. Ainsi, le score maximal $P_l^* = \max\{P_l\}$ permet de caractériser la matrice l^* correspondant au mieux au fragment X_S. Cette matrice va être donc légèrement modifiée pour ressembler un peu plus au fragment X_S. Chaque fréquence d'acides aminés f_{aa} en position k est modifiée.

- pour l'acide aminé o présenté à la position k du fragment X_S:

$$f_o^k \leftarrow \frac{f_o^k + \alpha}{1 + \alpha}$$

- pour les 19 autres types d'acides aminés à la même position :

$$f_{aa}^k \leftarrow \frac{f_{aa}^k}{1 + \alpha}$$

Cette transformation permet de conserver en chaque position une somme des fréquences

toujours égale à un. Le coefficient d'apprentissage α est égal à $\alpha_0/(1 + t/N_x)$, avec α_0 le taux initial d'apprentissage pris égal à 0,05, t représentant le nombre de fragments déjà vus et N_k, le nombre total de fragments associés au bloc protéique x. Le processus est itératif, l'ensemble des fragments est donc vu totalement à chaque cycle. Au bout d'un certain nombre de cycles, les fragments se focalisent sur une seule des f matrices. 5 cycles ont été utilisés dans cette apprentissage.

4.3.2 Construction des familles séquentielles

La figure 4.6 illustre le principe de la séparation en deux matrices du bloc protéique b en deux familles séquentielles. On peut observer sur les matrices normalisées en Z-scores une différence de localisation des sur- et des sous-représentations. Ceci se retrouve dans les zones de plus grande informativité obtenues par le profil KLd (cf. figure 4.7). Le Kld maximal est passé de 0,1 à 0,3. En n'observant que les valeurs supérieures à 0,08, on voit que pour la première famille séquentielle la zone d'intérêt se trouve dans l'intervalle -3 à +2, ainsi que pour les positions (-7) et (+4); pour le second, la zone est restreinte à l'intervalle [-2;+2]. Leurs modes aussi sont différents, respectivement en (-1) et (0).

En comparant les deux matrices associées, des différences nettes en composition en acides aminés sont visibles en la première et la seconde famille séquentielle, comme une sur-représentation en Alanine en position (-7), Acide Aspartique (-2), Proline en (-1), Histidine et Aspartate en (0), Proline en (+1) and Phénylalanine en (+6), ainsi que des sous-représentations en Lysine en position (-2), Glycine en (+1) et Cystéine en (+4). Les caractéristiques principales du bloc protéique b sont retrouvées dans ses deux familles séquentielles, comme la sous-représentation en Proline en position (+2).

Des essais pour tout les blocs ont été effectués en prenant un nombre de familles f compris entre 2 et 6. Les blocs divisés en familles séquentielles ont été choisis en prenant comme critère le taux de prédiction globale au premier rang, soit $\mathbf{Q}(1)$. Comme plusieurs matrices pour le même bloc protéique étaient utilisées, seule celle ayant le plus haut score est conservée pour la prédiction bayésienne. Un autre critère a été pris en compte, il s'agissait de rééquilibrer les taux de prédiction entre tous les blocs. Le tableau 4.1 récapitule le nombre de familles séquentielles conservées. Le taux de prédiction $\mathbf{Q}(1)$ est passé de 34,4 % à 40,7 % (gain de 6,3 %), avec les

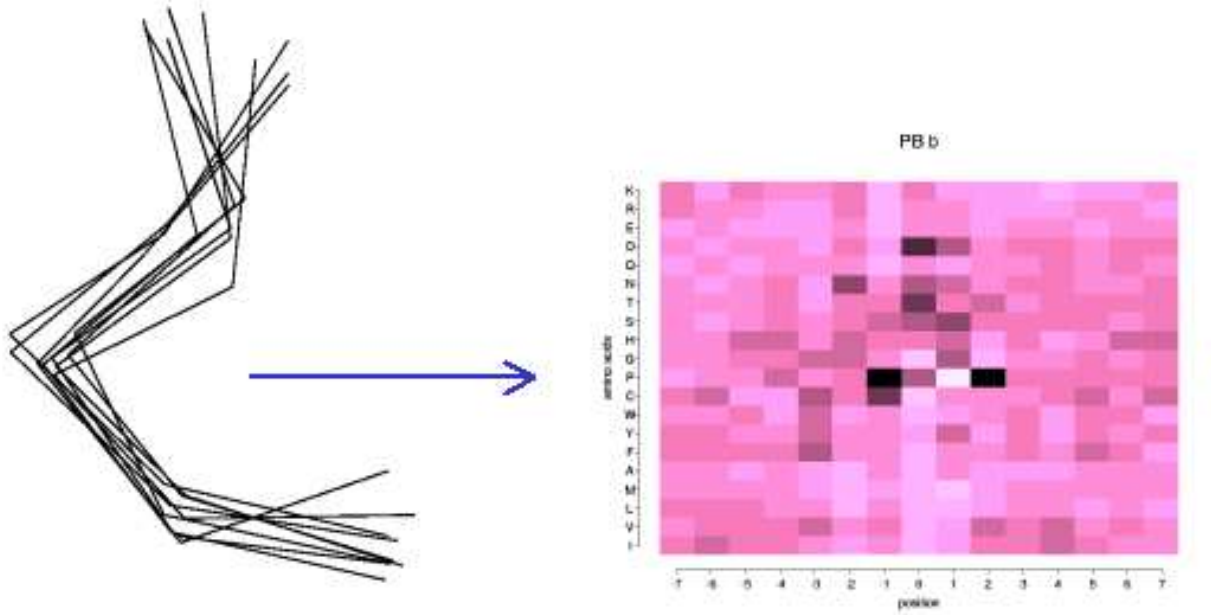


FIG. 4.6 – *Principe du découpage d'une matrice d'occurrence pour un bloc proté*
familles séquentielles.

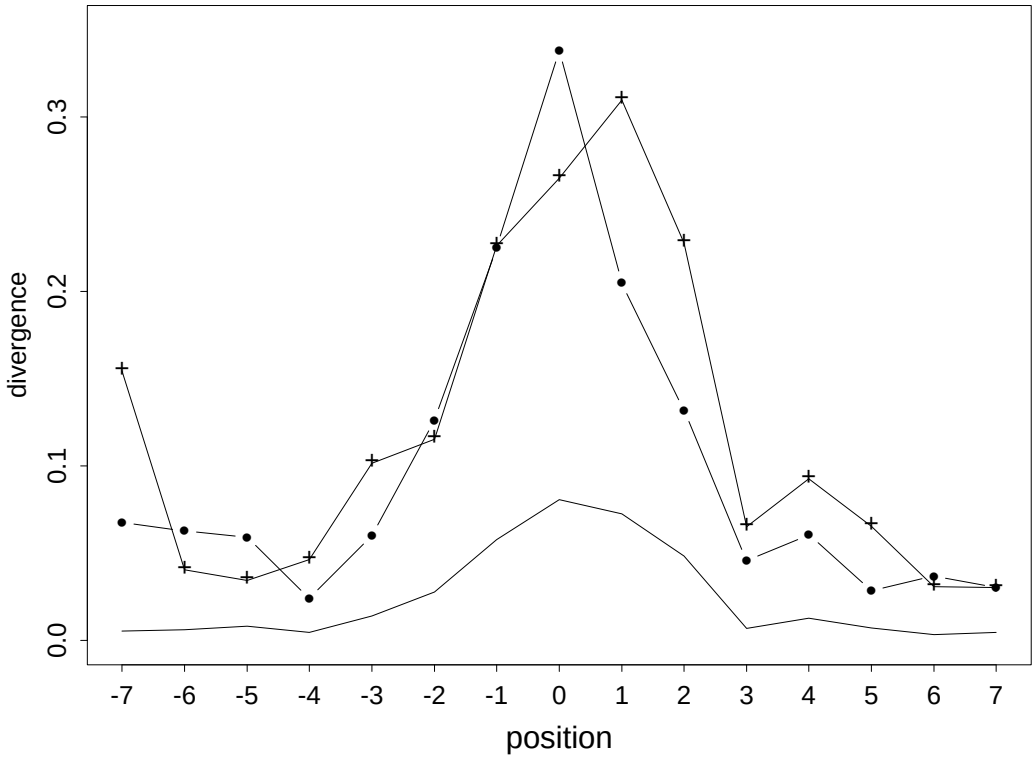


FIG. 4.7 – *Exemple de l'évolution du KLd dans le découpage en famille séquentielle pour le bloc protéique* b*, avec le profil KLd initial en ligne pleine et les profils des Kld des deux matrices issues de ce bloc en ligne avec des points et avec des ronds.*

bloc	nombre de familles séquentielles	taux de prédiction initial	taux de prédiction fam. séq.
a	1	60,3	53,5
b	2	13,1	27,0
c	2	26,3	32,9
d	3	27,6	34,8
e	1	43,1	35,9
f	2	32,2	36,2
g	1	31,4	35,1
h	1	52,1	42,7
i	1	42,1	41,0
j	1	57,5	47,2
k	1	41,4	35,2
l	1	39,2	32,1
m	6	36,1	50,8
n	1	56,0	44,7
o	1	55,8	45,8
p	1	33,5	33,9
Total	26	34,4	40,7

TAB. 4.1 – *Nombre de familles séquentielles par bloc, avec leur taux de prédiction pour l'approche initale et celle utilisant les familles séquentielles (fam. séq.).*

taux de prédiction initial et avec l'utilisation des familles séquentielles.

En résumé, ce sont les blocs les plus fréquents qui ont pu être subdivisés, la fréquence finale de leur famille séquentielle se rapproche d'ailleurs alors de la fréquence des autres blocs protéiques.

Par ailleurs, il a fallu vérifier que la création de ces nouvelles familles séquentielles construites sur le plan de la séquence n'ait pas eu de conséquence sur le plan de la structure. Aussi, pour chaque famille séquentielle, le vecteur d'observation moyen des 8 angles dièdres (cf. paragraphe 3.2) le caractérisant a été calculé à l'aide des fragments appartenant à chaque nouvelle matrice. Ces vecteurs ont été comparés au vecteur décrivant le bloc dont ils sont issus (cf. tableau 3.1) ainsi qu'à celui des autres familles séquentielles du même bloc. Un seul angle se trouve à plus de 3 degrés de différences. En conclusion, les familles séquentielles n'ont pas créé de "nouveau" bloc protéique.

4.3.3 Influence des familles séquentielles dans la prédiction

Avec les 26 matrices obtenues par l'utilisation des familles séquentielles, le taux de pédiction $\mathbf{Q}(1)^*$ est passé à 40,7 %, en conservant les deux solutions les plus probables ($\mathbf{Q}(2)^*$), le taux est de 57,5 %, soit 5,4 % de gain par rapport à l'approche Bayésienne simple, de même $\mathbf{Q}(4)^*$

= 75,8 % (gain de 4,4 %) et atteint 90,2 % pour $\mathbf{Q}(7)^*$, soit un gain d'un rang pour le même taux de probabilité.

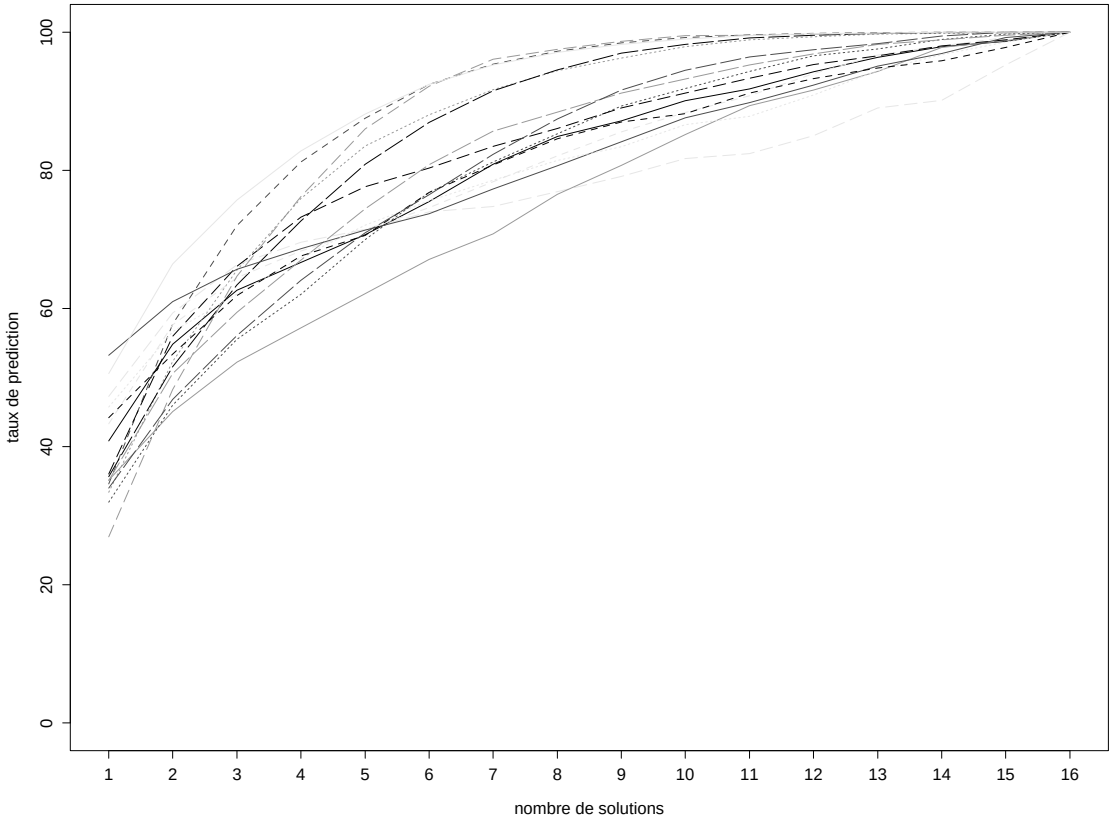


FIG. 4.8 – *Taux de prédiction pour chaque type de blocs protéiques en conservant de 1 à 16 solutions possibles avec utilisation des familles séquentielles*

La figure 4.8 montre, comme la figure 4.5 pour la prédiction bayésienne simple, l'évolution du taux de prédiction individuel des BPs en fonction du nombre de solutions conservées. L'effet de concentration des blocs dans un intervalle plus restreint est obtenu. L'écart entre le BP le meilleur et le plus mal prédit est passé de 47,2 % à 26,2 % avec une augmentation du taux de prédiction du BPb de 13,1 % à 27,0 % et une diminution de celui du BPa de 60,3 % à 53,2 %, ce dernier n'ayant pas été divisé.

La figure 4.9 montre la différence qui existe entre le taux de prédiction initial $\mathbf{Q}(1)$ indiqué en abscisse et la différence entre ce taux et $\mathbf{Q}(1)^*$ obtenu par les familles. Cette figure montre bien que le gain concerne la majorité des protéines, 95 % ont gagné en taux de prédiction. Maintenant 51,4 % des protéines ont un taux de prédiction supérieur à 40 % contre moins de 21 % auparavant. Ce gain n'est pas équivalent selon le type de protéine ainsi en moyenne les protéines tout-α ont un gain de 9,1 % (37,3% contre 46,4 %), les tout-β ont un gain plus faible

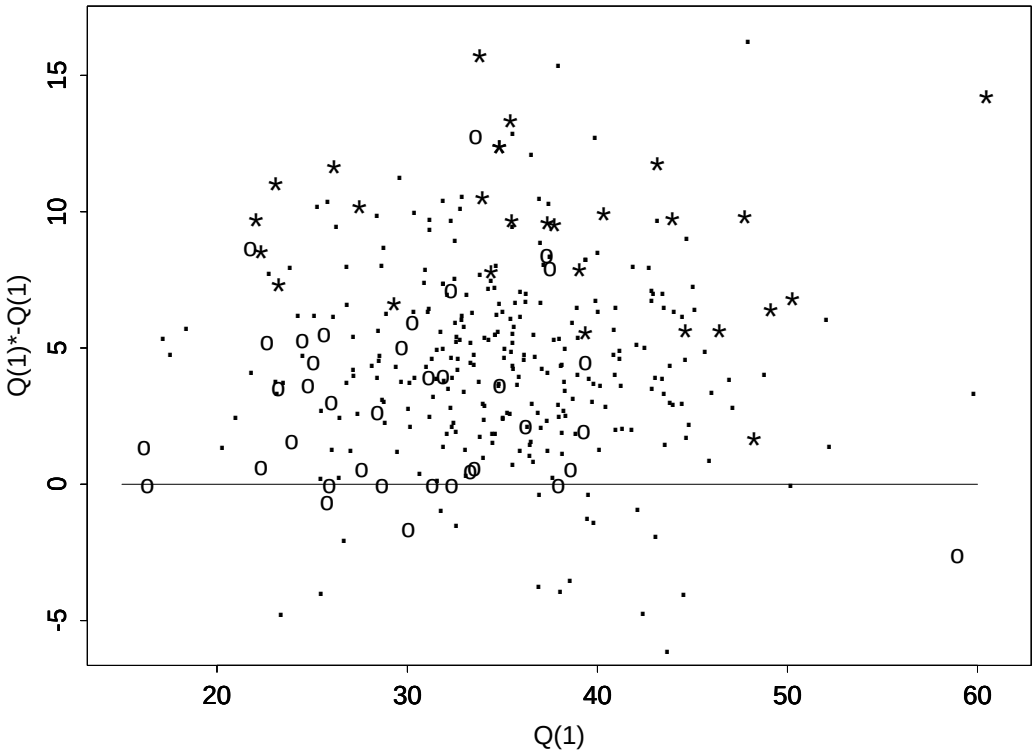


FIG. 4.9 – *Gain du taux de prédiction par l'utilisation des familles séquentielles, avec en abcisse le taux de prédiction* $\boldsymbol{Q}(1)$ *pour la méthode bayésienne simple et en ordonnée la différence entre ce taux* $\boldsymbol{Q}(1)$ *et celui obtenu avec les familles séquentielles* $\boldsymbol{Q}(1)^*$*. Les taux ont été donnés pour chaque protéine de la base de données: (*) les protéines tout-*α*, (o) les protéines tout-*β *et (.) les autres protéines; la classification suit la définition de Michie et collaborateurs [127].*

de 3%, (30,2 % contre 33,2 %), les $\alpha+\beta$ ont un gain de 4.9% (35,7 % - 40,6 %), et 4,8% pour les non-classifiées (33, 9% - 38,7%).

Globalement, on observe une augmentation des taux de prédiction pour les blocs ayant plusieurs familles séquentielles et en contre-partie une diminution pour les autres. Toutefois, cette diminution est faible. Il aurait été simple d'augmenter artificiellement les taux de prédiction des blocs les plus fréquents en les sub-divisant encore plus, jouant ainsi sur l'effet de l'importance numérique du bloc. Mais, alors les blocs moins fréquents auraient vu leur taux individuel décroître rapidement en dessous des 10 %, le taux global devenant par ailleurs largement supérieur à celui obtenu ici. La figure 4.8 montre que le nombre de familles séquentielles obtenues est raisonnable.

Nous pouvons noter que ce ne sont pas les blocs des structures répétitives les mieux prédits; le bloc protéique m (hélice α) a un taux de 50,6 % et le bloc protéique d (feuillet β) de 34,6 %, le PB a, une entrée en feuillet β, atteint lui un taux de 53,2 %.

4.4 Stratégies de prédiction

4.4.1 Le "nombre équivalent de blocs" (N_{eq}): un indice de confiance de prédiction

Après avoir prédit à partir de la séquence le (ou les) bloc protéique(s) le(s) plus probable(s) en appliquant le principe des familles séquentielles (*n séquences - 1 repliement local*). Nous avons décidé d'introduire un concept flou *1 séquence - n repliements* qui prend en compte le fait qu'une séquence peut être associée à plusieurs types de repliements, donc plusieurs types de blocs protéiques. En observant les résultats de la prédiction, le bloc réel est souvent le plus probable, mais il est surtout fort souvent parmi les plus probables. En conséquence, j'ai essayé de définir des stratégies pour sélectionner le nombre optimal de blocs r à prendre en compte en chaque site pour avoir un taux de prédiction donné.

Dans la suite de ce chapitre, deux types de stratégies distinctes vont être utilisées. Elles se basent toutes sur deux l'entropie de Shannon et sur le fait qu'une grande homogénéité de scores R_k en un site donné veut dire que l'informativité de la séquence X_S doit être faible. La prédiction associée localement est alors peu fiable. Inversement, un score élevé pour le bloc

protéique le plus probable doit être associé à un bon taux de prédiction. Dans le premier cas, il faudra choisir un nombre élevé de blocs, alors que dans le second, il en faudra moins. Pour quantifier cette incertitude, une entropie H a été calculée sur les scores R_k. Ces scores ont été dans un premier temps renormalisés en probabilités S_k :

$$S_k = \frac{R_k}{\sum_{l=1}^{l=B} R_l}$$

avec l d'ecrivant l'ensemble des blocs protéiques, avec $B = 16$ dans notre cas.

L'expression de l'entropie est alors :

$$H = -\sum_{l=1}^{l=B} S_l \ln(S_l)$$

Ensuite, l'entropie H est transformée en nombre équivalent de blocs noté N_{eq}:

$$N_{eq} = exp[H]$$

Cette quantité varie entre 1 quand un seul bloc est prédit, et, B quand les B blocs sont équiprobables. Les sites ayant un N_{eq} variant entre 1 et un N_{eq}^g (g allant de 1 à 8) ont été extraits de la base de données et le pourcentage de bonne prédiction Q_r a été ainsi calculé pour r rangs avec r variant entre 1 et 6, le bloc réel étant trouvé parmi les r rangs conservés. Les blocs sont tout d'abord classés par ordre de score décroissant.

De cette distribution, associée avec un intervalle de N_{eq} donné, nous déterminons le nombre rang optimal r pour assurer un taux de prédiction fixé. Cette étape a été effectuée pour tous les rangs possibles, allant de 1 à B.

Deux stratégies différentes ont ainsi été définies à partir des observations précédentes :

(i) *une approche globale* qui consiste à définir un nombre variable de blocs à conserver en chaque site s pour atteindre un taux global de prédiction Q_g fixé au préalable. Dans cette optique le nombre de blocs protéiques conservés varie le long de la séquence.

(ii) *une approche locale* qui tend à trouver les sites permettant pour un nombre fixé r de blocs conservés d'obtenir un taux de prédiction Q_l, lui aussi fixé au préalable. Dans cette approche, la prédiction est limitée à certaines régions des séquences protéiques.

Nous verrons donc dans une première partie, l'influence des familles séquentielles sur le N_{eq}, puis un exemple de prédiction sur un fragment de protéine pour voir l'évolution du N_{eq}, puis enfin successivement les deux stratégies.

4.4.2 Influence des familles séquentielles sur le N_{eq}

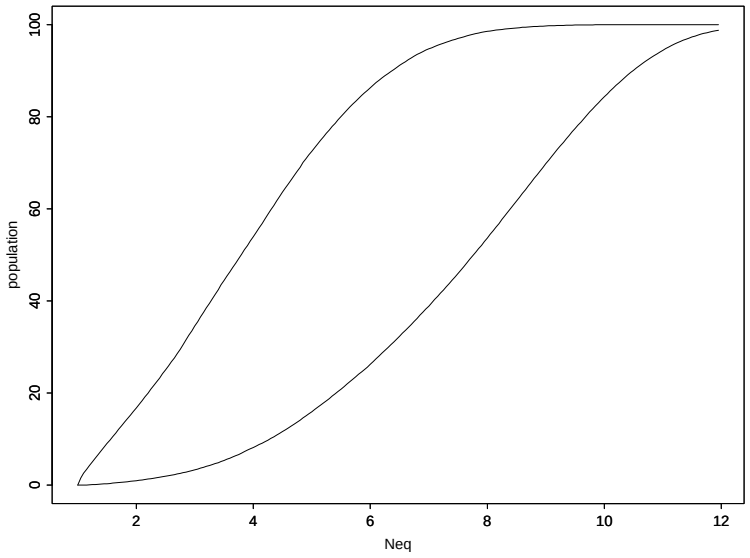


FIG. 4.10 – *Evolution du N_{eq} entre l'approche Bayésienne simple (courbe du bas) et l'utilisation des familles séquentielles (courbe supérieure).*

Par ailleurs, l'utilisation des familles séquentielles entraîne, du fait de la spécialisation séquentielle des blocs, une diminution du N_{eq} moyen comme reportée sur la figure 4.10 qui met en parallèle l'évolution du N_{eq} en fonction du nombre d'observation de la base de données, pour la prédiction par l'approche bayésienne simple et avec l'utilisation des familles séquentielles. Pour les blocs ayant plusieurs familles séquentielles, seul le score le plus élevé est conservé, donc le N_{eq} diminue; ceci change fortement l'allure de la courbe obtenue. Le N_{eq} moyen est passé de 7,6 à 4,2.

4.4.3 Exemple de prédiction

Pour mieux expliciter l'intérêt des stratégies, l'exemple suivant montre un exemple de la prédiction et le rôle du N_{eq}. Le tableau 4.2 donne les prédictions des 18 premières positions de la protéine de conjugaison à l'ubiquitine (cf. paragraphe 3.3.2.4), avec la fenêtre de 15 résidus correspondant au 5 C_α du bloc et aux 5 résidus présents de part et d'autre de cette fenêtre.

sequence			bloc réel	N_{eq}
gauche	centrale	droite		
DMSTP	**ARKLM**	RDFKR	*l*	2,74
MSTPA	**RKLMR**	DFKRL	*m*	2,43
STPAR	**KLMRD**	FKRLQ	*m*	2,06
TPARK	**LMRDF**	KRLQQ	*m*	2,63
PARKL	**MRDFK**	RLQQD	*m*	2,48
ARKLM	**RDFKR**	LQQDP	*m*	3,78
RKLMR	**DFKRL**	QQDPP	*m*	2,92
KLMRD	**FKRLQ**	QDPPA	*m*	3,49
LMRDF	**KRLQQ**	DPPAG	*m*	6,32
MRDFK	**RLQQD**	PPAGI	*m*	8,61
RDFKR	**LQQDP**	PAGIA	*m*	4,82
DFKRL	**QQDPP**	AGIAG	*c*	2,55
FKRLQ	**QDPPA**	GIAGA	*c*	3,10
KRLQQ	**DPPAG**	IAGAG	*e*	5,45
RLQQD	**PPAGI**	AGAGI	*h*	5,34
LQQDP	**PAGIA**	GAGIS	*i*	4,97
QQDPP	**AGIAG**	AGISG	*a*	4,75
QDPPA	**GIAGA**	GISGA	*c*	6,75

TAB. 4.2 – *Exemple de prédiction de la partie N-terminale de la protéine de conj*
séquence prédite avec la partie centrale représentant le bloc structural et son en
réel, le N_{eq} *et les trois blocs les plus probables classés par ordre décroissant. Le b*

Cette partie N-terminale est composée d'une hélice α formée par 10 blocs protéiques m suivi par une boucle de 7 blocs qui mène à un feuillet β. Cet exemple est basé sur l'utilisation des familles séquentielles précédemment décrite. Chaque ligne correspond à une séquence, par exemple, la cinquième fenêtre centrée sur MRDFK est assignée au bloc protéique m. Les trois premières solutions ont été ordonnées suivant leur score de prédiction R_k, pour BP m, BP f et BP b, leurs scores respectifs étant de 22,13, 1,25 et 0,40.

Ainsi, le premier score indique que la probabilité du bloc m est 22,13 fois plus élevée que celle d'avoir ce bloc de façon purement aléatoire. En cette position, la prédiction est correcte. Les scores élevés des premières positions sont justifiés par la présence de résidus Leucine, Méthionine, Arginine, Lysine, Aspartate et Leucine en position (-3), (-2), (-1), (+2), (+3) et (+4). De même, le BP f est classé en seconde position du fait de la prsence de l'Aspartate en position centrale de la fenêtre. En ne conservant que les premiers rangs, 10 blocs protéiques sont correctement prédits sur 18. Sur l'ensemble des protéines, le taux de prédiction Q(1)* est de 40,8%. Sans tenir compte des familles séquentielles, il était de 30,4%, soit un gain de plus de 10%. Classiquement, les taux de prédiction ne sont calculés que pour les solutions optimales. Mais, en observant, les solutions des trois premiers rangs, 17 des 18 blocs y sont. La position erronée correspond à une fin d'hélices α qui possède une composition inhabituelle en acides aminés, KRLQQDPPA en [-4;+4].

Aussi, au lieu de ne prendre en compte que les premiers rangs, une approche pertinente revient à examiner les taux de prédiction Q(r) pour un rang donné r. Le N_{eq} permet de quantifier cette dispersion parmi les scores. Ainsi dans la première partie de l'hélice α, le N_{eq} varie entre 2,06 et 3,78; il est ainsi corrélé avec une bonne prédiction. Inversement, à la fin de l'hélice α, la probabilité de trouver le bloc protéique réel décroît alors que le N_{eq} augmente au-delà de 4,82. Les sites sont de moins en moins informatifs. Des N_{eq} intermédiaires sont observés pour les 7 derniers résidus, le nombre de rang à conserver est alors de 2.

Cet exemple montre l'intérêt des stratégies de prédiction basées sur un nombre variable de blocs sélectionnés par site.

4.4.4 Stratégie globale

En utilisant la base de données, nous avons établi la relation qui existe entre la probabilité de trouver le bloc réel parmi les r blocs les plus probables pour un N_{eq} donné. Cette démarche permet d'obtenir la figure 4.11 qui met en relation le taux de prédiction en fonction du N_{eq} et du nombre de solutions conservées. Cet exemple utilise les familles séquentielles.

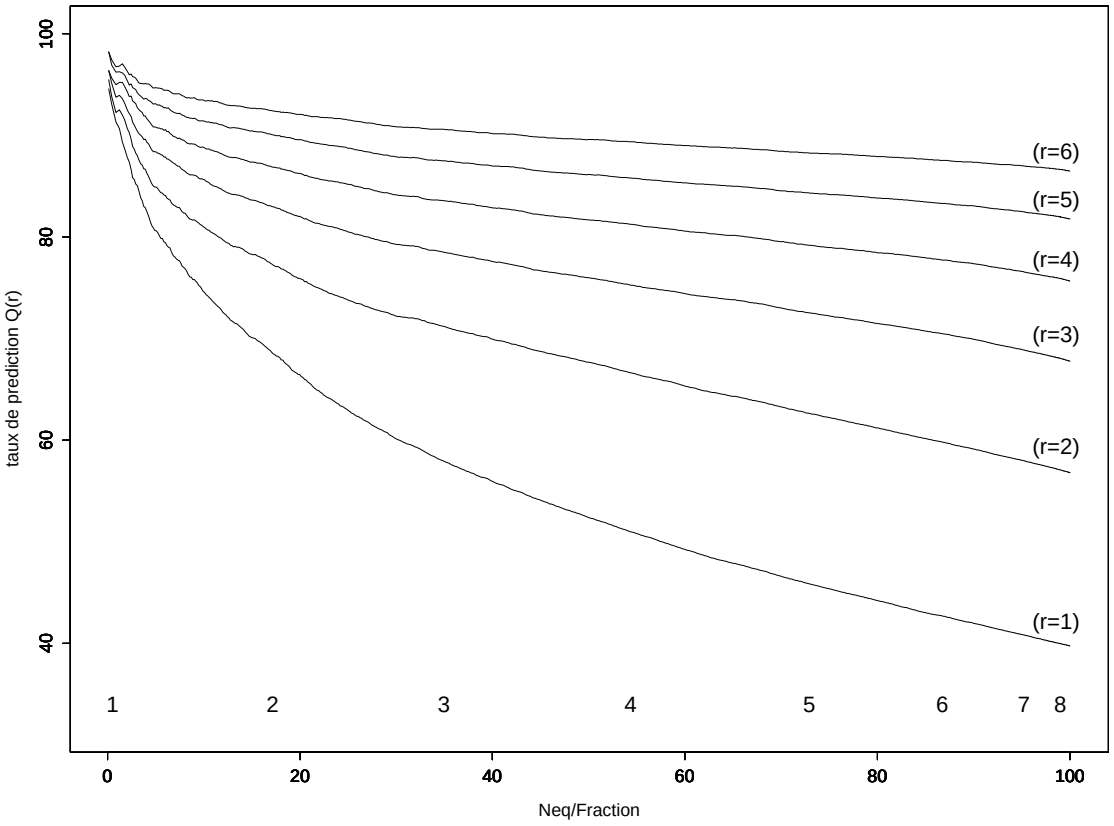


FIG. 4.11 – *Taux de prédiction* $\boldsymbol{Q}(r)$ *associé à chaque taux de Neq pour r variant entre 1 et 6.*

La distribution du taux de prédiction associé à chaque N_{eq} a été calculée. Pour chaque taux de prédiction Q_g, le nombre de rangs, (i.e. nombre de blocs) à conserver pour atteindre Q_g en fonction du N_{eq} a été déterminé. Par exemple, pour un N_{eq} inférieur à 6,32, il faut sélectionner les 3 PBs les plus probables pour avoir un taux de bonne prédiction de 70 %.

La figure 4.12 montre le résultat de cette stratégie pour la protéine *1aak* (cf. paragraphes 3.3.2.4 et 4.2). Le profil des N_{eq} (figure 4.12a) montre la variation de cet indice entre 1,06 et 9,79. La figure 4.12b donne en chaque site le rang véritable du bloc dans la prédiction. 77,8% des blocs réels sont parmi les 3 solutions les plus probables. Certaines zones de la protéine nécessite de conserver un grand nombre PBs, comme les deux boucles reliant les deux feuillets

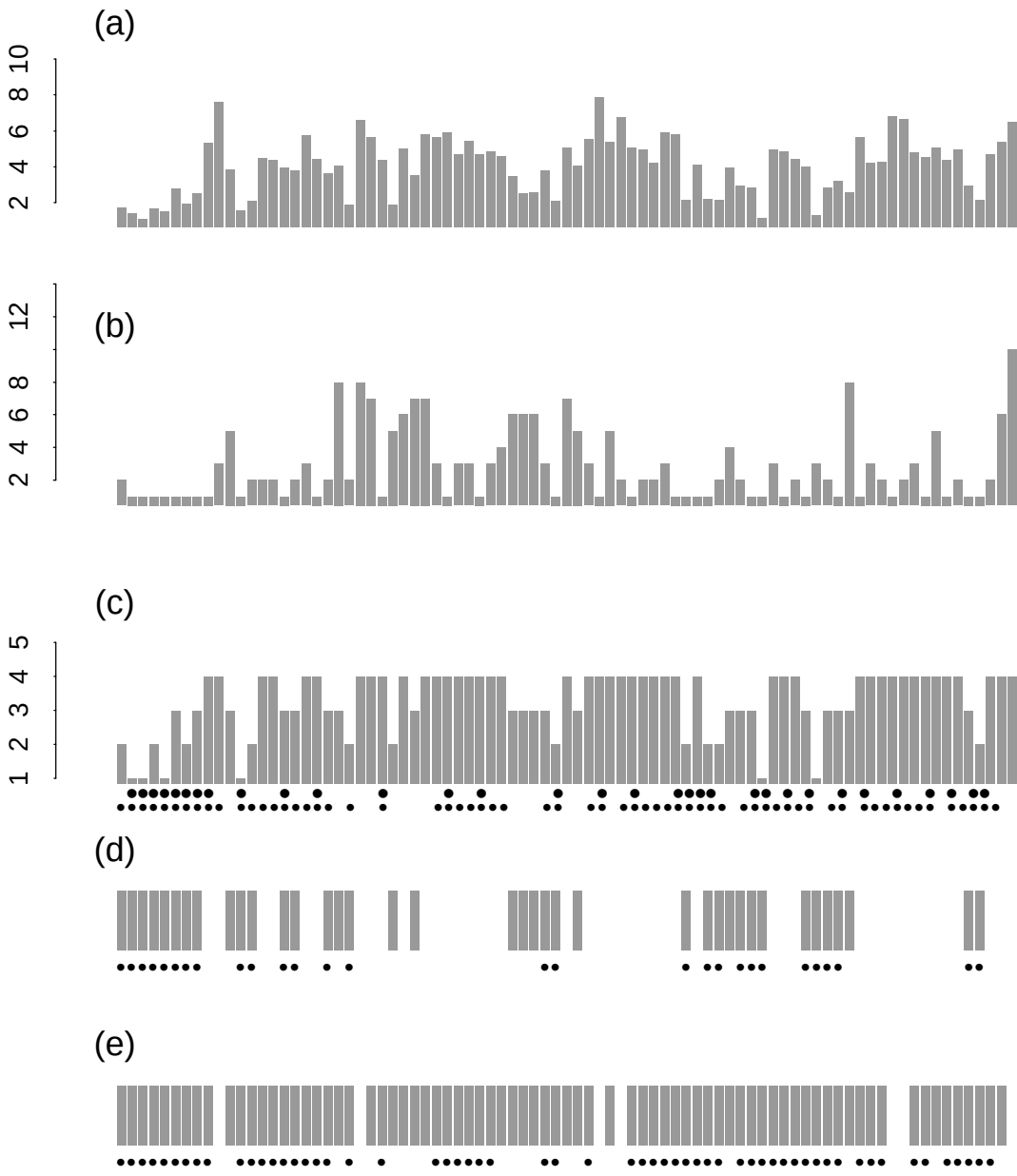


FIG. 4.12 – *Application des 2 stratégies à la protéine de liaison à l'ubiquitine ave*
du bloc réel dans les solutions, (c) taux global de prédiction Q_g de 75 %, avec
positions où le bloc réel est au premier rang, les points plus importants sont les
zone conservée pour la stratégie locale avec un taux de prédiction Q_l de 75 %, en
% et r = *3 rangs.*

β (positions 22 à 46), et la large boucle (positions 82 à 90) contenant une petite hélice α.

La figure 4.12c montre le nombre de blocs protéiques devant être sélectionné pour avoir un taux global de prédiction Q_g de 75%. Les séries de points en-dessous définissent les sites où le bloc réel est trouvé au premier rang, et parmi les rangs sélectionnés. Le nombre maximal de blocs protéiques est de 4. Le taux de prédiction au premier rang est de 40,7%; pour Q_g=75%, de 1 à 4 blocs protéiques par site sont sélectionnés (8, 17, 37 et 72 sites respectivement). Les structures répétitives et les blocs proches de ces dernières (cf. figure 3.10) sont bien délimitées. Comme attendu, les boucles sont plus difficilement prédictibles. En observant les deux séries de points, il est net que les zones possédant des blocs bien prédits au premier rang sont les plus aisées à prédire avec un nombre restreint de blocs.

Cette stratégie amène à un excès de blocs en chaque position. Cependant, en contrepartie le taux de prédiction est toujours assuré.

4.4.5 Stratégie locale

La seconde stratégie diffère de la précédente qui prenait un nombre variable de blocs protéiques sur l'ensemble de la protéine, en utilisant un nombre constant de blocs sur une partie de la protéine. Ainsi, un taux de prédiction Q_l est garanti pour r blocs choisis. La figure 4.11 montre chaque valeur de N_{eq} et pour un nombre r variant de 1 à 5. Pour calculer ces courbes, les sites compris entre 1 et chaque valeur de N_{eq} ont été conservés et le taux de prédiction associé calculé. Cet étude a été répétée depuis r=1 (juste le premier rang, le plus probable des blocs) jusqu'au 6 premiers rangs inclus.

Pour donner un exemple, si l'on désire avoir 70 % des sites le N_{eq} doit être inférieur à 4,8. En sélectionnant 1 (de même 2, 3 et 4) rang(s), le taux de prédiction associé est de 46,8 % (de même 63,4 % 73,1 % et 79,6 %). De la même manière pour un taux de prédiction donné de 80 %, le N_{eq} est de 1,28 et 5,5 % des sites seront sélectionnés avec un seul rang; ils passent respectivement à 1,6 et 11,5 % pour les deux premiers rangs, 2,6 et 26,9 % pour les trois premiers, et, 4,6 et 66,4 % pour les quatres premiers.

La figure 4.12d est un exemple de cette stratégie appliquée à la protéine *1aak*, les zones prises en compte représentent un taux de prédiction Q_l de 75%, en conservant 3 rangs. Le N_{eq} correspondant est alors inférieur à 5,11. 62 ont été sélectionnés, les points en dessous montrent

les 49 positions où le bloc réel se trouve parmi les blocs sélectionnés. Le taux final de prédiction est de 79 % pour 46,3 % des sites de la protéine pris en compte. En comparant avec la précédente approche, il est clair que prendre 3 blocs de manière fixe est un excès. De la même manière avec $r = 4$ et $Q_l = 75$ %, 52 % des résidus de la protéine sont alors utilisés.

La figure 4.12e montre la même stratégie pour Q_l=70 % et $r = 3$ rangs. En utilisant un N_{eq} maximal de 6,32, 122 sites, soit 91 % des sites de la protéine ont été sélectionnés et 95 de ces sites possèdent le bloc réel parmi ceux choisis soit un taux de prédiction de 77,9 %.

Ainsi, cette stratégie permet de localiser les sites les plus prédicitibles, cependant une recherche préalable doit être menée quand au nombre r de rangs qui doit être sélectionné. Par exemple, pour un taux de prédiction de Q_l=70 %, la proportion des sites sélectionnés augmente fortement avec un passage de r =2 à 3 rangs. (une augmentation de 49 %). Pour de future application de ces stratégies, comme dans des méthodes *ab initio*, le choix du nombre de blocs sélectionnés par site pose un certain problème : augmenter le nombre de rangs conservé r permet une prise en compte d'une plus grande partie des sites, mais aussi induit une combinatoire plus complexe pour reconstruire un modèle moléculaire.

4.5 Conclusion

La figure 4.13 récapitule l'ensemble du processus bayésien mis en place. Dans un premier temps, les séquences ont été directement utilisées en donnant un taux de prédiction plus que convenable de 34,4 % pour 16 états possibles. La définition des familles séquentielles (*1 bloc* $\rightarrow$ *n séquences*), permet à la fois un gain global de prédiction et une homogénéisation des taux de prédiction des blocs en conservant une homogénéité structurale des blocs protéiques. L'augmentation du taux de prédiction est significatif avec un passage de 34,4 à 40,7 %.

La succession d'une certaine série d'acides aminés n'obligent pas un seul type de repliements [186], mais notre approche permet de voir que cette succession induit un certain type de repliement qui peut être particulièrement bien caractérisé. La disposition préférentielle des blocs réels parmi les blocs les plus probables a permis l'élaboration du concept (*1 séquence* $\rightarrow$ *n blocs*) avec l'utilisation d'un indice de confiance le N_{eq} qui permet de bien localiser les zones les plus probables.

Ces recherches ont permis la mise au point de deux stratégies distinctes pour rechercher les

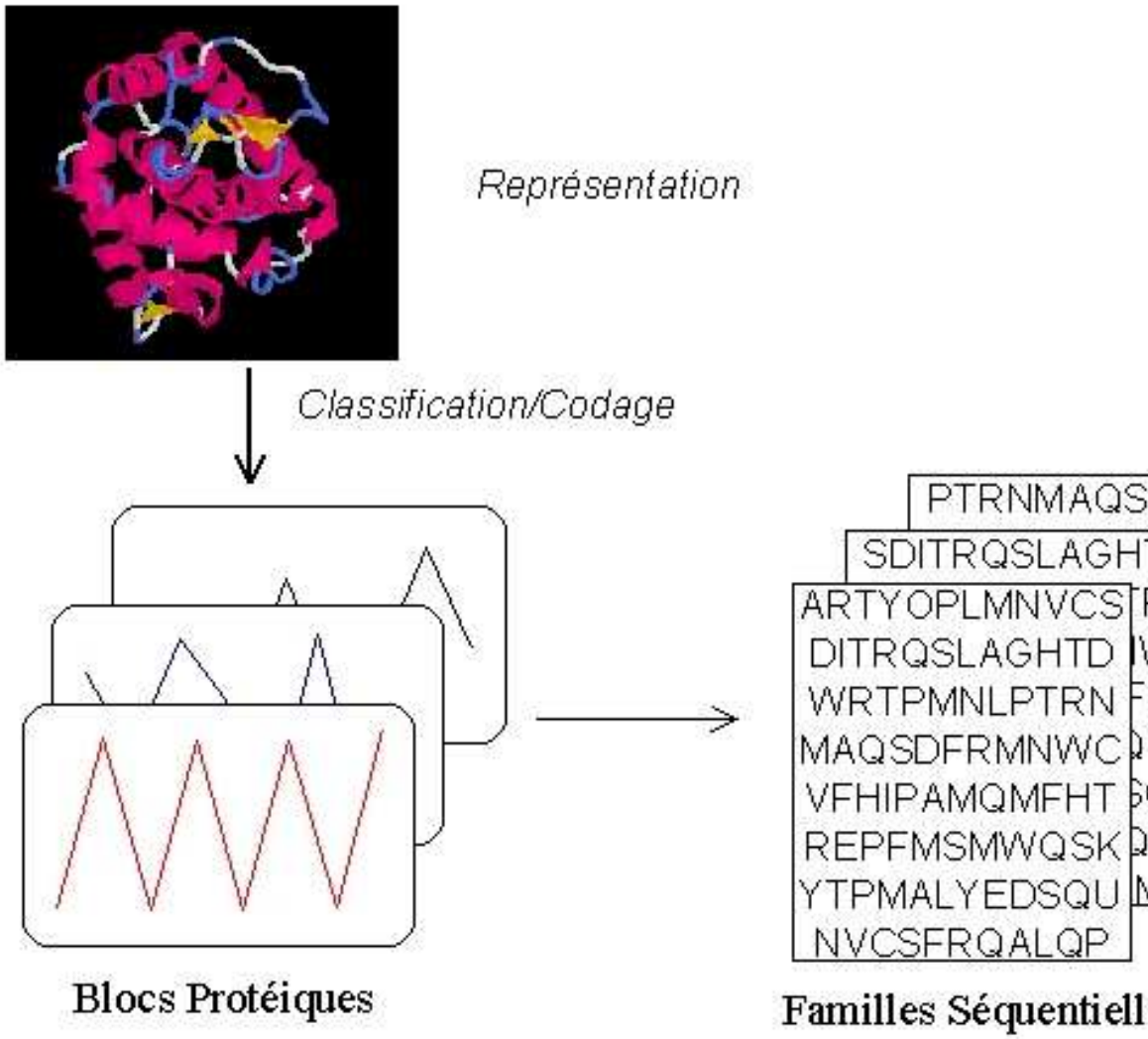


FIG. 4.13 – *Schéma récapitulatif de la prédiction bayésienne, aprés la réalisation à leur bloc et cette information est directement utilisée par la prédiction.*

zones et/ou le nombre de blocs à conserver pour aboutir à un taux de prédiction donné. La première stratégie donne un nombre variable de blocs en chaque résidu suivant le N_{eq} local, la seconde avec un nombre fixe de blocs donne un taux de prédiction garanti pour un nombre de zones prédite limitée. Ces deux stratégies donnent toujours un nombre trop élevé de blocs à conserver, il faudra y remédier dans un proche futur. L'utilisation de méthode classique type réseau de neurone artificiel et/ou alignements de séquence aurait surement donné des résultats quantitativement supérieurs. Toutefois, ces méthodes sont des techniques qui ne permettent en aucun cas de comprendre les tenant et les aboutissant de manière globale. Développant un nouvel alphabet largement plus complexe que les structures secondaires classiques, il aurait été regrettable de ne pas comprendre les spécificités de chacun des blocs et de passer directement à un apprentissage type "boite noire". Maintenant que cette première étape a été effectuée, avec le Pr. Hazout, nous développons une approche de type réseau neuronal artificiel dont nous attendons beaucoup.

Chapitre 5

Dépendance entre les blocs structuraux protéiques

5.1 Objectif

Après avoir développé une méthode d'apprentissage de blocs protéiques (cf. paragraphe 3, *"Apprentissage de la structure locale du squelette protéique"*), nous avons utilisé ces blocs dans une méthode bayésienne de prédiction locale de ces blocs à partir de la séquence (cf. paragraphe 4, *"Prédiction de la structure locale en blocs protéiques"*). Conserver seize PBs au final, permet d'obtenir des BPs purement boucles ainsi que plusieurs types d'extrémités N- et C-terminales des structures périodiques.

Toutefois, cette approche comme beaucoup d'autres approches locales ne ne prend pas implicitement en compte le problème de la continuité implicite de la structure protéique. Ce problème a été déjà soulevé avec la prédiction des structures secondaires [210]. Le réseau neuronal PHD [166] le résout en traitant les résultats de sa prédiction par un second réseau. Actuellement, seules les Chaînes de Markov Cachées [149] utilise directement cette information, comme HMMSTR, pour les structures secondaires, développé par le groupe de Baker [20] et pour les alphabets structuraux par Camproux et collaborateurs [23, 24].

Pour prendre en compte cette continuité, nous avons décidé d'analyser les séries de BPs ayant les occurrences les plus fréquentes, par une approche qui présente certaine similitude avec le travail de Fetrow et collaborateurs [53]. Les séries de blocs ont été établies en tenant en compte des transitions préférentielles. Ces successions de blocs les plus fréquentes se suivent de manière particulièrement régulière. Les chevauchements sont nombreux.

Nous proposons donc de construire un réseau simple formé par la succession de ces mots.

Ce travail est à rapprocher d'autres descriptions des structures par des graphes pour trouver des fragments protéique structuralement proches [170], des groupes de chaînes latérales [101], des domaines protéiques [201], mais surtout des "topologies proches" dans les protéines, en d'autres termes des fragments protéiques de tailles variables ayant des repliements locaux similaires [106, 191], nous regarderons ensuite si la construction d'un tel réseau permet une bonne conservation structurale des fragments protéiques et enfin si la présence de certains acides aminés est dépendante des voies différentes du réseau. Cette approche est intéressante car au lieu de se focaliser sur des zones de longueur constante entre des structures secondaires [203, 204], des zones plus ou moins étendues seront obtenues. Ainsi, les séries de blocs permettent d'observer et d'analyser de manière fine la structure 3D sans avoir les limites liées au choix de longueur fixe.

5.2 Conception du graphe

5.2.1 Construction du réseau

Nous nous sommes intéressés aux mots (i.e. *succession*) de 5 blocs protéiques, ce qui représente 9 C_α soit la taille moyenne d'un feuillet β avec entrée et sortie complète. La base de données est celle utilisée auparavant (342 protéines ayant moins de 25% d'identité de séquence).

L'ensemble des structures protéiques a été recodée en BPs selon la règle du *RMSda* minimal. Le réseau que l'on désire construire est un graphe orienté. Un graphe **G** correspond à un ensemble **T** de nœuds ("vertices") et d'un ensemble **E** de segments ("edges") qui les relient. Le graphe est orienté si chaque segment possède une seule direction, **G**(V,E). Dans notre étude, chaque nœud de l'ensemble **V** est caractérisé par un bloc protéique, et chaque lien orienté une transition entre deux blocs protéiques. L'objectif est de caractériser le graphe orienté des séries de blocs protéiques les plus fréquemment observés dans une base de structures protéiques. Pour réaliser cet objectif, les séries de 5 blocs protéiques les plus fréquents (fréquence supérieure à 150 observations soit 0,18 % de la base étudiée) ont été sélectionnés, puis en se basant sur un principe de séquentialité, le graphe orienté a été construit. Un motif de 5 blocs est représenté par un sous-graphe orienté. Par exemple, le motif *mnopac* est décrit par le sous-graphe m $\rightarrow$ n $\rightarrow$ o $\rightarrow$ p $\rightarrow$ a $\rightarrow$c.

Le principe de séquentialité consiste à trouver dans la liste des motifs ceux dont les 4 derniers blocs d'une série se retrouvent dans les 4 premières d'une autre série, ($a_1 \rightarrow a_2 \rightarrow a_3 \rightarrow a_4 \rightarrow a_5$) et ($a_2 \rightarrow a_3 \rightarrow a_4 \rightarrow a_5 \rightarrow a_6$) devient ($a_1 \rightarrow a_2 \rightarrow a_3 \rightarrow a_4 \rightarrow a_5 \rightarrow a_6$). Par exemple, le motif *mnopa* est suivi par le motif *nopac* par continuité. Par superposition des sous-graphes, le graphe est étendu. Le sous-graphe devient dans ce cas m→ n→ o→ p→ a→c. Le graphe peut présenter des bifurcations, une série étant suivie par plusieurs séries distinctes, ainsi, le motif *mnopa* est suivi par deux motifs *nopac* et *nopaf* dans des proportions respectives de 68 et 32 %. Les structures secondaires répétitives hélices et feuillets sont considérées chacune comme un seul nœud retournant sur lui même ainsi la succession *fklmmmmmmmmmmnop* se traduira par la succession des nœuds *fkl(m*)nop*. La répétitivité des blocs sera notée par une étoile. Cette méthode va nous permettre de bien mettre en relief des transitions à plus longue distance.

5.2.2 Qualité du réseau

Pour s'assurer de la stabilité structurale du réseau, un ensemble de N motifs qui permettent de recouvrir le graphe totalement ont été extraits. Les *RMSds* des fragments extraits de la base de données associés à ces mots ont été calculés pour évaluer leurs sensibilités structurales.

5.3 Résultats

5.3.1 Le réseau

Sur l'ensemble des 342 structures protéiques recodées, 72 motifs de taille 5 ont été obtenus. Leurs fréquences varient entre 17,3 % (14729 observations) et 0,18 % (152 observations). Les *RMsd* ont été calculés pour l'ensemble de ces motifs, appellés mots structuraux (les valeurs et la description de ses mots sont disponibles sur le site web de l'Equipe de Bioinformatique Génomique et Moléculaire). Les motifs les plus fréquents sont associés aux formes des structures secondaires (*mmmmm* et *ddddd*). Les 14 premiers ont servi à mettre en place le réseau et 44 des 58 suivants ont servi ŀe compléter. La figure 5.1 montre le réseau obtenu. Ce réseau est ainsi composé de 31 nœuds et utilise 15 des 16 BPs. Les couleurs représentent les occurrences observées, ainsi l'entrée principale en hélice α *fkl* (nœuds *08-09-10*) vers le BP *m* (nœud *01*) a une sortie principale *nop* (nœuds *02-03-04*) et deux sorties plus courtes (et moins observées) *pcc*

(nœuds *20-21-22*) et *c* (*23*). La partie entourant le site du BP *d* (nœud *07*) est plus complexe, l'entrée s'effectue par un seul nœud *c* (nœud *06*), celui-ci étant compris dans une succession *nopac* (nœuds *02-03-04-05-06*) ou *kbc* (nœuds *31-27-06*), sa sortie la plus importante est la série *dehiacd* (nœuds *07-25-16-17-18-19-07*). La seconde sortie du feuillet la plus importante est la succession *dfkl* (nœuds *07-08-09-10*). Ensuite la sortie plus complexe de *df* (nœuds *07-13*) allant vers les sites *b* (nœud *14*) et *k* (nœud *31*).

Des parties moins observées ont était mises comme le dédoublement du BP *e* (nœud *26-15*) en amont du site *h* (nœud *16*), la sortie en *j* (nœud *25*) de ce dernier, la suite alternative de *fkl* (nœuds *08-09-10*) vers *pc* (nœuds *11-12*) ou les deux passages de *dfk* (nœuds *07-13-31*) vers *bc* (nœuds *27-06*) ou *op* (nœuds *03-04*). Le réseau final obtenu est visible dans un plan 2D sans aucune intersection des flèches. Le caractère discontinu du réseau doit être noté, ainsi tous les fragments conservés sont ceux qui peuvent s'inscrire dans le réseau, c'est à dire s'ils ont 5 sites consécutifs. La figure 5.2 donne les fréquences de début et de fin des motifs observés sur le réseau, cad la fréquence relative de chaque nœud pour être le premier ou le dernier résidu d'un fragment protéique compris dans le graphe. Seul le site *c* (nœud *06*) situé juste avant le bloc *d* (feuillet β, nœud *07*) possède des taux largement supérieurs aux autres sites. Ce fait s'explique à la fois par des différences de tailles et de formes de structures associées aux feuillets β et surtout au principe même du graphe où 3 des 4 terminaisons sont des nœuds correspondant au PB *c*: en fin d'hélice α nœuds *22* et *23* des séries *mc* et *mpcc* (en haut à gauche de la figure 5.1) et le nœud *12* de la série *lpc* (en bas du graphe).

5.3.2 Exemple

Pour mieux appréhender cette notion de réseau non continu, la figure 5.3 représente la partie centrale de l'endoribonuclease du virus sarcome aviaire (code PDB: 1vsd) avec sa structure protéique recodée en blocs protéiques et en-dessous les différentes parties du réseau alors utilisées. Trois zones appelées 'fragment' se trouvent incluses dans le réseau. La figure 5.3 montre cet exemple avec en haut la séquence en BPs, et, en-dessous les trois fragments avec les zones du graphe correspondantes. Le fragment **I** est composé de 7 blocs protéiques, c'est un court feuillet *bcddfbf* (partie supérieure droite du réseau). Le fragment suivant **II** se superpose au dernier bloc protéique puis repart vers une forme α entrant par la successions de blocs *fkl* et en res-

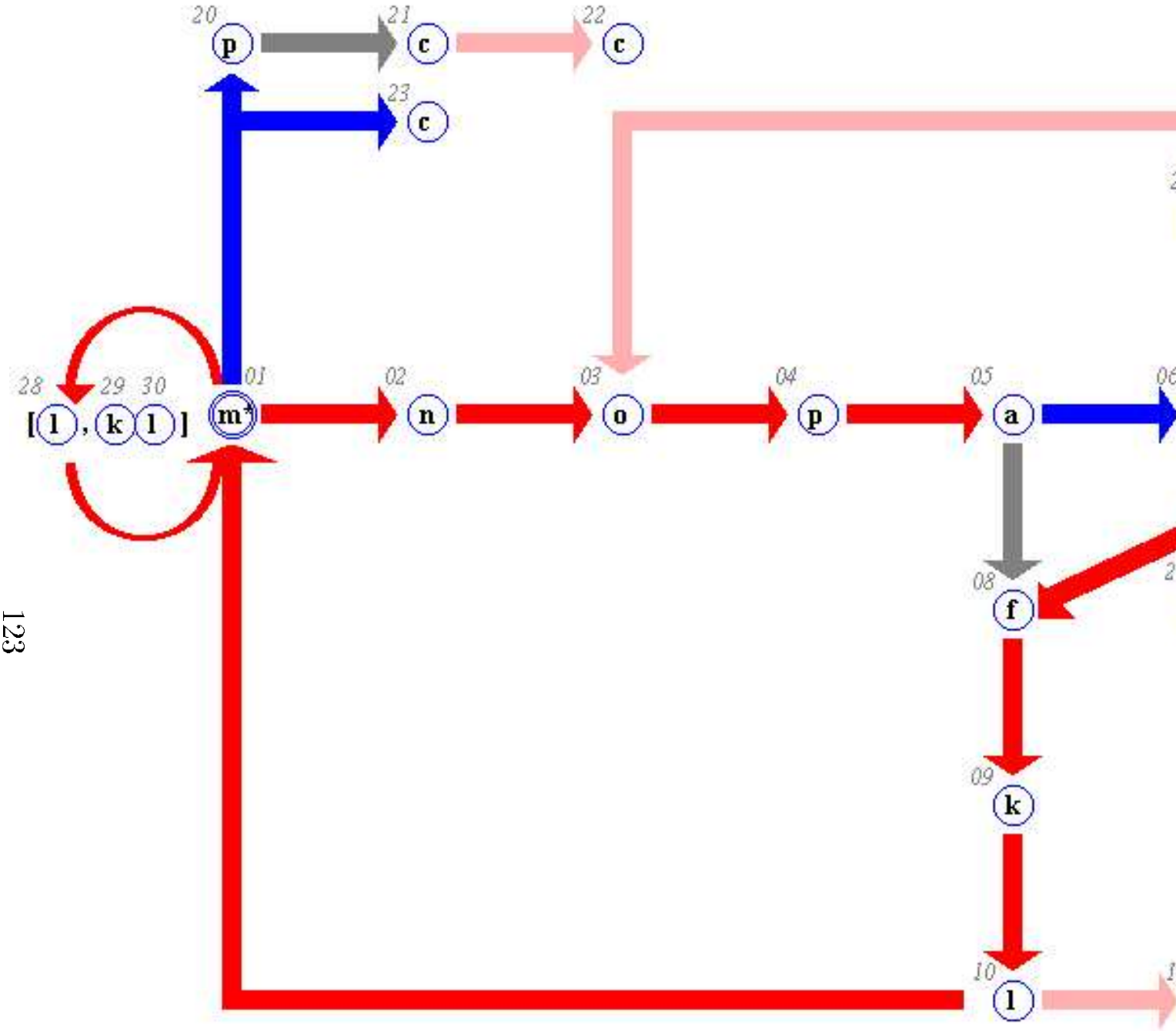


FIG. 5.1 – *Le réseau obtenu contient 31 nœuds, avec au-dessus le numéro du r rouge, des occurrences > 800 observations (10 %), en bleu entre 500 (6 %) et 80 rose en-dessous de 300 (3,5 %). Les sites m* et b* sont des répétitions des blocs site l (nœud 28) et les sites kl (nœuds 29-30) représente des successions mlm et r mpcc, mc, lpc, le site j de la suite hj et le site b de bf sont des terminaisons du g*

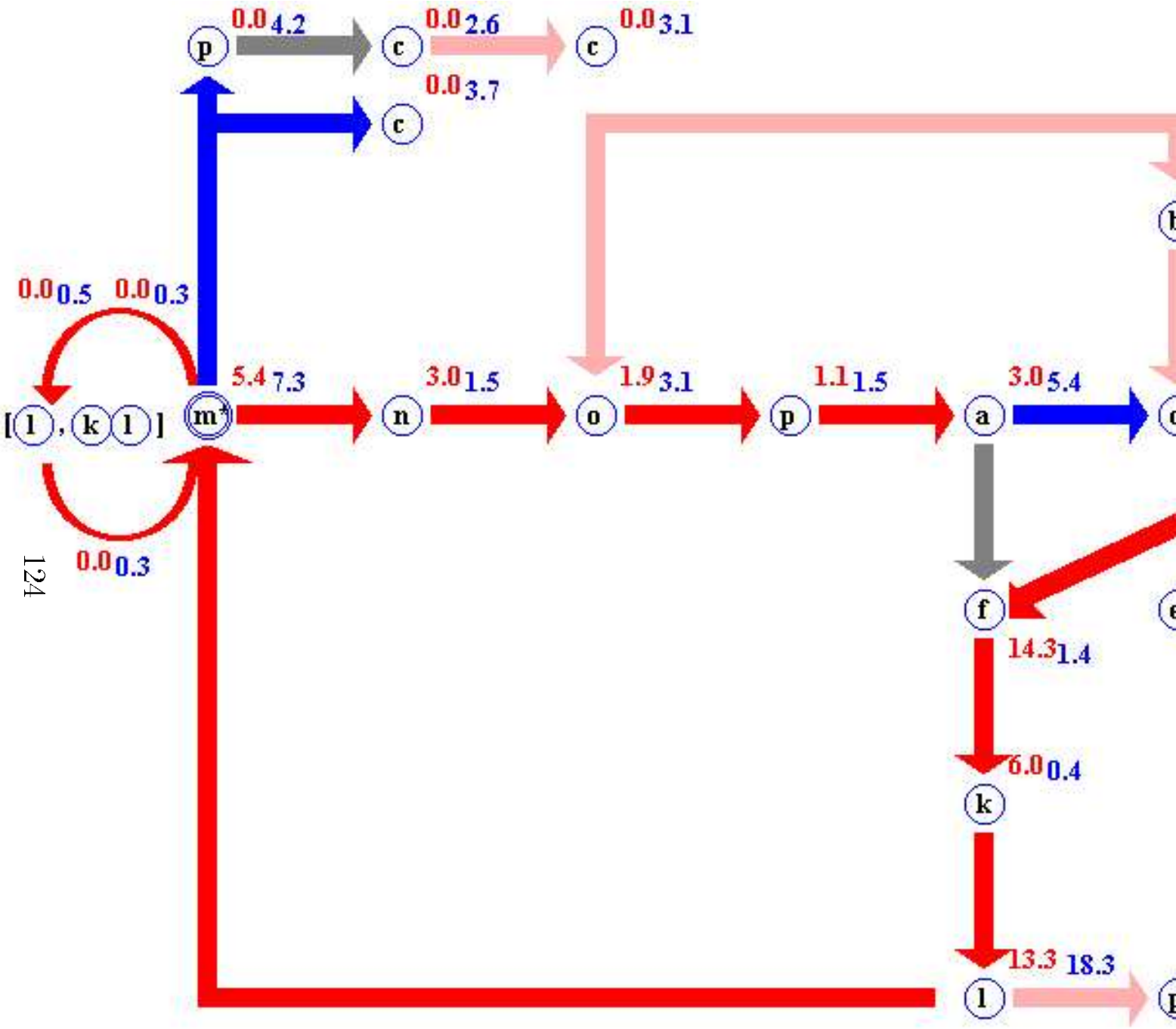


FIG. 5.2 – *Les pourcentages de début et de fin pour les 31 nœuds du réseau. L*
rouge, et celles de fin juste en-dessous des précédentes en bleues.

sortant par la série de blocs *nopacde* (partie inférieure gauche du réseau, puis partie centrale). Ensuite une zone composée par une série de blocs *bjghehpap* est non observée parmi les motifs les plus fréquents et donc n'a pas de contre-partie dans le réseau. Enfin le fragment **III** est une longue hélice qui est composé de 19 blocs protéiques *m* consécutifs (partie inférieure gauche du réseau).

5.3.3 Analyse du réseau

Ayant recodé l'ensemble des protéines à l'aide de l'alphabet structural, plusieurs questions se posent quand à la pertinence et à la stabilité structurale du réseau : (1) Quelle est la part des parties non-répétitives dans le réseau? (2) Est-ce que les fragments protéiques contenus au même site du réseau sont bien structuralement identiques?

Pour répondre à la première question, l'ensemble des protéines a été suivi sur le graphe, comme pour l'exemple (cf. figure 5.3) et les taux des résidus appartenant au réseau a été calculé. Avec cette métrique (un minimum de 5 BPs consécutifs compris dans le réseau) et le réseau proposé, plus de 83 % des acides aminés des protéines est inclus dans le graphe, mais le point le plus intéressant n'est pas que les structures répétitives secondaires soient bien retrouvées (96,14 % pour les hélices et 89,11 % pour les feuillets) mais que 70,58 % des boucles soient aussi présentes.

Le premier point ayant été validé, il faut se pencher sur le second car, en effet, l'alphabet structural a été construit pour approximer localement la structure tridimensionnelle sur 5 C_α et le réseau utilise des successions chevauchantes de 5 BPs (soit 9 C_α), il convient donc de vérifier que ces approximations locales n'ont pas généré des fragments protéiques trop distincts.

5.3.4 Extraction des mots et stabilité 3D

17 mots de tailles 4 à 7 BPs (i.e. 8 à 11 C_α) ont été extraits du graphe, ils permettent de le recouvrir intégralement. Le *RMSd* moyen a été calculé sur l'ensemble des fragments protéiques correspondants dans la base de données. Celui des mots structuraux a été effectué préalablement. Le tableau 5.1 récapitule les mots observés et les *RMSds* associés. Dans une seconde étape, nous avons regardé s'il n'était pas possible dans ces mots de discerner des sous-familles structurales, cad des repliements distincts, mais associés aux mêmes successions de

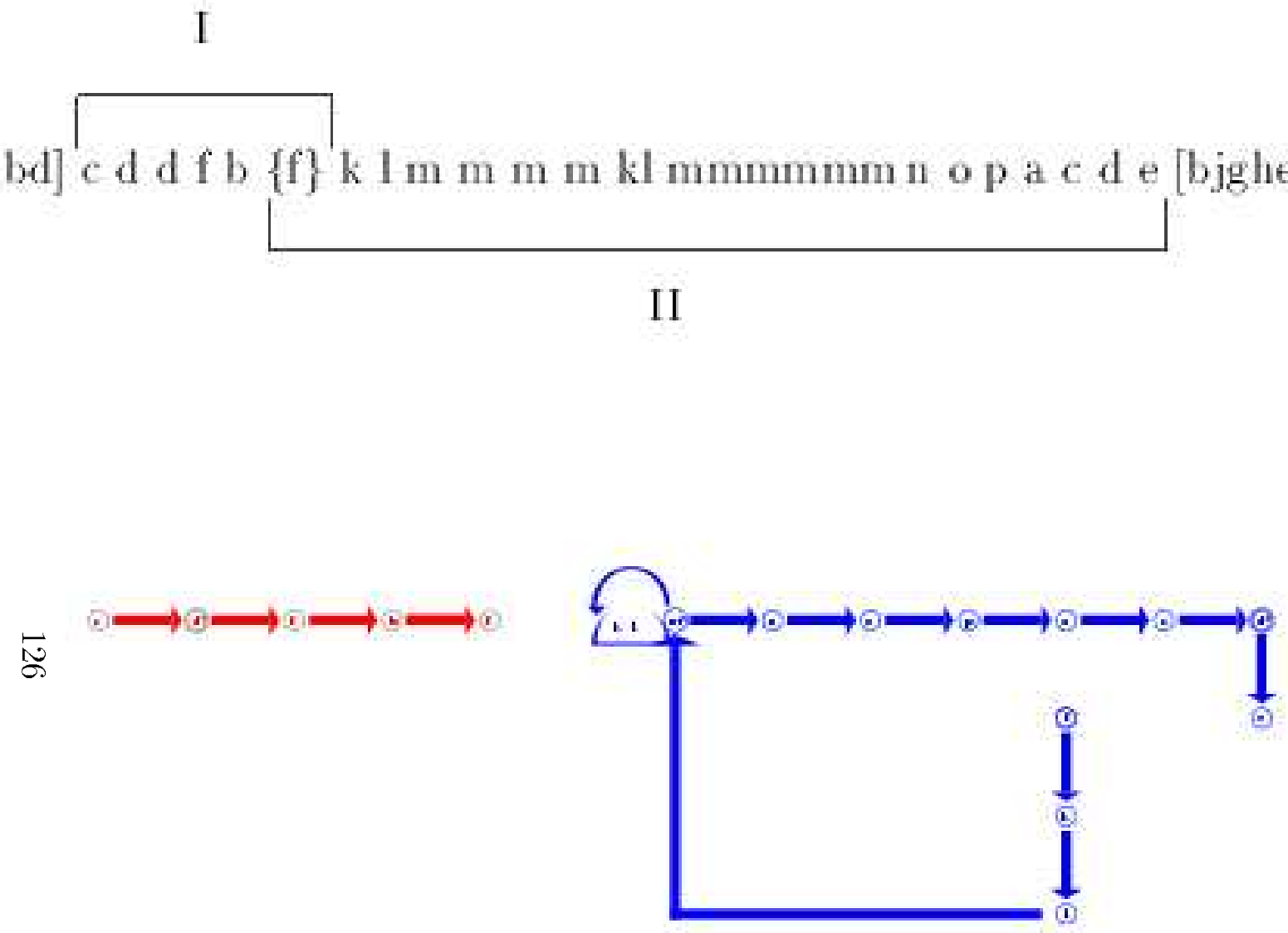


FIG. 5.3 – *Exemple de recodage d'une séquence protéique (virus sarcome aviaire protéique recodée en blocs et en dessous les différentes parties du réseau alors im*

num	motif	BPs	C_α	*RMSd*m *Å*	nœuds
1	mnopacd	7	11	1,94	01 - 02 - 03 - 04 - 05 - 06 - 07
2	mnopa	5	9	1,76	01 - 02 - 03 - 04 - 05
3	ehiac	5	9	2,68	15 - 16 - 17 - 18 - 19
4	dehiacd	7	11	3,07	07 - 15 - 16 - 17 - 18 - 19 - 07
5	dfbf	4	8	1,74	07 - 13 - 14 - 24
6	dfkopa	6	10	1,37	07 - 13 - 31 - 03 - 04 - 05
7	fklm	4	8	1,69	08 - 09 - 10 - 01
8	afkl	4	8	1,81	05 - 08 - 09 - 10
9	dfbd	4	8	2,02	07 - 13 - 14 - 07
10	mklmm	5	9	1,49	01 - 29 - 30 - 01 - 01
11	mlmm	4	8	1,42	01 - 28 - 01 - 01
12	mpcc	4	8	1,74	01 - 20 - 21 - 22
13	fklpc	5	9	1,93	08 - 09 - 10 - 11 -12
14	eehiac	6	10	1,34	26 - 15 - 16 - 17 - 18 - 19
15	dehj	4	8	1,98	07 - 15 - 16 - 25
16	dfkbc	5	9	1,94	07 - 13 - 31 - 27 - 06
17	dfklm	5	9	2,09	07 - 08 - 09 - 10 - 01

TAB. 5.1 – *Pour chaque motif est donné la succession des blocs protéiques compris dans le motif testé, le nombre de Blocs Protéiques les composant et le* RMSd *moyen (en Å) associé.*

BPs.

5.3.4.1 Test de l'homogénéite structurale au sens du *RMSd*

La superposition des C_α des fragments montre une constance remarquable avec un *RMSd* moyen pour la plupart des "mots" aux alentours de 2 Å. La figure 5.4 illustre le cas de la succession de blocs *mnopacd* qui possède un *RMSd* de 1,94 Å pour une longueur de 11 C_α (partie gauche et centrale du réseau, nœuds *01-02-03-04-05-06-07*). Ce ne sont pas obligatoirement des motifs liés à des hélices α qui sont le plus conservés, ainsi *dfkopa* (nœuds *07-13-31-03-04-05*) et *dfbf* (nœuds *07-13-14-24*) ont des *RMSd* de, 1,37 et 1,74 Å, respectivement.

Afin d'évaluer l'homogénéité du groupe de structures associé à un mot, les *RMSd*s entre les couples de fragments ayant été calculés, une matrice de distance a été constituée et a servi lors d'un regroupement par une méthode hiérarchique (paquage *hclust* du logiciel *S-Plus*). Dans chaque arbre obtenu par cette méthode 2 et/ou 3 groupes ont été formés, pour voir, s'ils n'existaient pas dans certaines séries des sous-familles structurales discernables. Ces groupes de fragments ont été donc de nouveaux superposés. Ils s'avérent qu'ils ne sont pas mieux

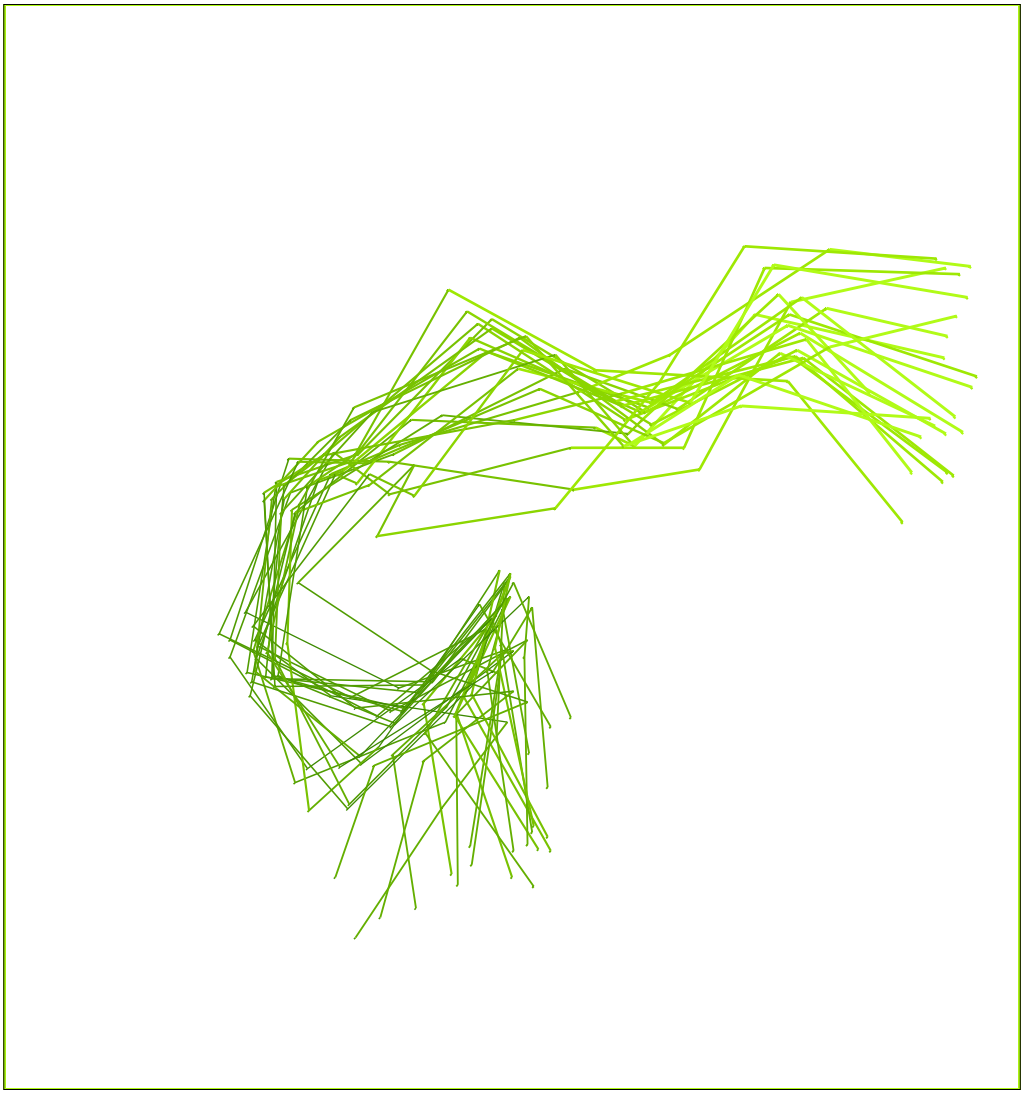


FIG. 5.4 – *Superposition de fragments représentant la succession de blocs mnopacd.*

définis que les motifs globaux et qui ne donnent pas des *RMSd* moyens meilleurs. Seul le motif "long" *dehiacd* (nœuds 07-15-16-17-18-19-07) possède une flexibilité qui semble importante; ses extrémités N- et C- terminales parraissent plus mobiles (distance de 13 Å+/-5 Å), toutefois le repliement interne étant caractéristique, aucun découpage utilisant le *RMSd* n'est possible.

Ainsi, les motifs pris dans le réseau ont une stabilité structurale correcte sans être aussi bonne que dans le cas des BPs seuls (0,46 Å à 1,04 Å pour une longueur de 5 C_α). Le repliement observé apparaît globalement semblable pour des fragments associés à un motif donné.

5.3.4.2 Test de l'homogénéite structurale au sens du RMSda

Les blocs n'ont pas été définis sur la base du *RMSd* mais du *RMSda* (*root mean square on angular values*) qui par définition est beaucoup plus sensible. Les *RMSda* moyens ont été calculés pour les même couples de fragments que pour le *RMSd*. Leurs valeurs moyennes étaient toutes proches de 30^o (avec une distribution gaussienne classique). Il ont donc une approximation exactement similaires à ceux des BPs seuls, ce qui est intéressant car la même précision est conservée pour un nombre d'angles passe de 8 à 14 et jusqu'à 20 angles, la spécificité semble donc assez grande.

Toutefois, il est vrai que le *RMSda* est une mesure plus difficile à apprécier. Une simple superposition des angles permet un aspect plus "visuel", la recherche d'angles précis ayant une variablité plus grande est arbitraire, mais pour la majorité des mots un seul type de signal est visible. Les séries *dehiacd* (nœuds 07 -15-16-17-18-19-07) et *mlmm* (nœuds 01-28-01-01) sont les seules séries ayant un angle qui montre une plus grande variabilité. Cependant en utilisant cet unique critère pour rechercher de possibles sous-familles, les résultats ne sont pas améliorés. En conclusion, sur un plan structural, le réseau permet d'approximer la structure 3D locale de manière assez correcte et ceci même pour des longueurs assez importantes.

5.3.5 Relation avec les acides aminés

D'un point de vue composition en acides aminés, la distribution des Z-scores en certains sites est ainsi beaucoup plus prononcée que pour les BPs seuls. Il a fallu tenir compte du fait qu'un bloc protéique est présent en plusieurs sites et donc que certains sites nettement moins observés que le BP seul ne sont pas statistiquement représentatifs. Certains BPs sont peu influencés d'un

point de vue compositionnel par leur environnement local, d'autres le sont fortement. La figure 5.5 montre un exemple les différentes répartitions en acides aminés pour les nœuds du réseau correspondant aux blocs protéiques *f*, *k*, *l*. Ainsi les nœuds *b*, *14* de la suite *df*$\underline{b}$*f* et *27* de *fk*$\underline{b}$*c*, ont peu de différence entre eux, le premier représentant 4/5 des sites *b* du graphe. L'absence de Proline dans le second nœud, alors qu'elle est fortement sur-representée pour le BP *b*, est compensée par une présence de Glycine, absente elle du BP *b*. Le cas des nœuds *e* est assez similaire pour le nœud *15* de la suite *d*$\underline{e}$*h* (89 %) et *15* de $\underline{e}$*eh* (11 %) avec peu de différence avec le BP *e* original, sauf pour la Proline et la Glycine totalement absente du nœud *15* de la série *eeh*. Les nœuds *c 06* de la série *a*$\underline{c}$*d* (68 % des *c* du graphe) et *21* de *mp*$\underline{c}$ (13 %) ont peu de différence sauf pour des hydrophobes moins présents dans le dernier cas. Les sites *k 09* de la série *f*$\underline{k}$*l* (75 %) et *29* de *m*$\underline{k}$*l* (18 %) ont peu de différence avec le BP *k* : une absence de Thréonine dans le premier et une sur-représentation de Sérine dans le second. Le principal nœud *l 10* de la série *kl*$\underline{m}$ (85 %) n'apporte que peu de spécificité avec juste une légère augmentation du nombre d'Alanine et une sous-représentation accrue de Glycine par rapport au BP *l*.

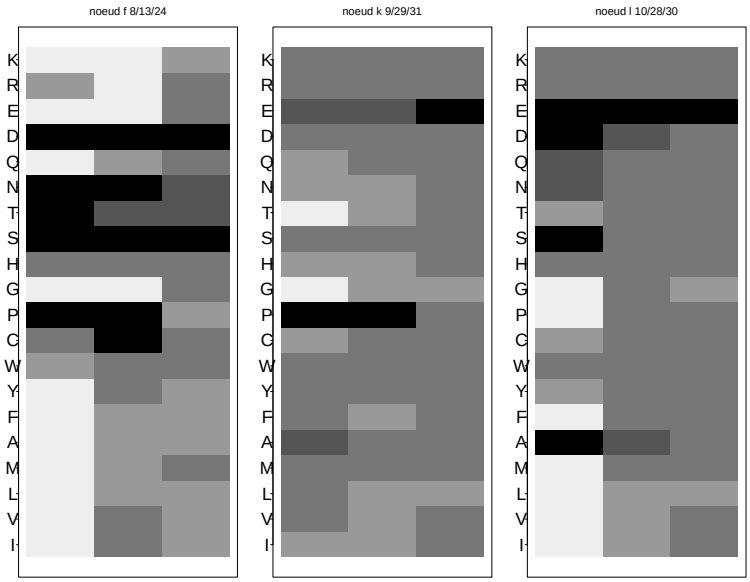


FIG. 5.5 – *Comparaison entre les distributions en acides aminés normalisés en Z-scores des nœuds correspondants aux blocs protéiques f (8, 13 et 24), k (9, 22 et 31) et l (10, 28 et 30).*

Les sites *p* des nœuds *04* de la série *op*$\underline{a}$ (65 %) et *20* de *mp*$\underline{c}$ (28 %) montrent des caractéristiques assez éloignées. Le premier est très proche de la composition du BP *p*, avec juste une accentuation de sous-représentations pour la Méthionine et le Tryptophane. Le second se différencie fortement par une présence d'acides aminés Isoleucine, Leucine, Alanine, Trypto-

phane et Cystéine inexistants dans le BPp, ainsi qu'une sur-représentation accrue de Lysine. Les nœuds *f 08* et *13* montrent dans les suites *fklm* (56 %) et *cdf* (42 %) une forte dissemblance en acides aminés. Le premier est assez proche du BP *f* seul, le second nettement moins avec une fréquence plus élevée en Isoleucine, Valine, Tyrosine, Tryptophane et Glutamine. Enfin, les nœuds *a 05* et *18* dans les suites *opa* (65 %) et *hia* (35 %) sont totalement antagonistes, le premier assez proche du BP *a* montre une surreprésentation en Isoleucine, Valine et Phénylalanine plus importante que le BP *a*, le second lui ne possède presque plus aucun acide aminé Isoleucine, Valine, Leucine, Méthionine, Tyrosine, Thréonine, Sérine, Glutamate, Lysine et Arginine et montre une forte propension en Glycine et Aspartate.

Longueur des protéines n'influence en rien le taux de résidus compris dans le graphe par protéines qui est assez constant. La longueur moyenne des fragments inclus dans le graphe n'en dépend d'ailleurs pas.

5.3.6 Les boucles

Quelques faits classiques, caractérisant les coudes, peuvent être notés. 50 % des coudes de type I de la PDB sont représentés par un bloc *k* et toujours à proximité de blocs de type *m*, ce qui se retrouve dans la partie inférieure gauche du graphe par les successions *fklm* (nœuds *08-09-10-01*) ou *mklm* (nœuds *01-29-30-01*). De même 80 % des coudes de type II se termine par *ia*, et plus de 50 % sont présents dans le réseau dans la succession *ehia* (nœuds *15-16-17-18*). Le coude de type III est comme le coude de type I toujours à proximité du bloc *m*, ainsi 80 % de ces structures sont caractérisées par la succession *lm* (nœuds *10-01/29-01*) et plus de 60 % par *klm* (nœuds *09-10-01/29-30-01*). Le graphe permet donc de retouver ces coudes classiques, mais sans se limiter à une définition trop stricte.

5.3.7 En dehors

Une grande partie de la base de données est comprise dans le réseau : que reste-t-il en dehors ? En analysant les doublets non compris dans le réseau, deux principaux en ressortent: les paires *bd* et *fb*. Les BPs situés aux extrémités N- et C-terminales complétant ces paires sont fortement variées. Le triplet *fbd* est compris dans le réseau" (partie droite du graphe avec les nœuds *13-14-07*), mais se retrouve à 413 reprises hors du réseau. Cette non-*prise en compte* dans le

réseau est dûe à la recherche effectuée : une taille minimale de 5 blocs consécutifs sur le réseau.

Il faut noter l'absence totale du bloc g qui est caractéristique de boucles de grandes tailles, plus variables. Il faudra donc noter que l'établissement d'un réseau prenant en compte un maximum de chemins a depuis été finalisé avec des chemins de fréquence plus faible, ainsi que l'utilisation de ces motifs répétés dans la prédiction de la structure à partir de la séquence.

5.4 Conclusion

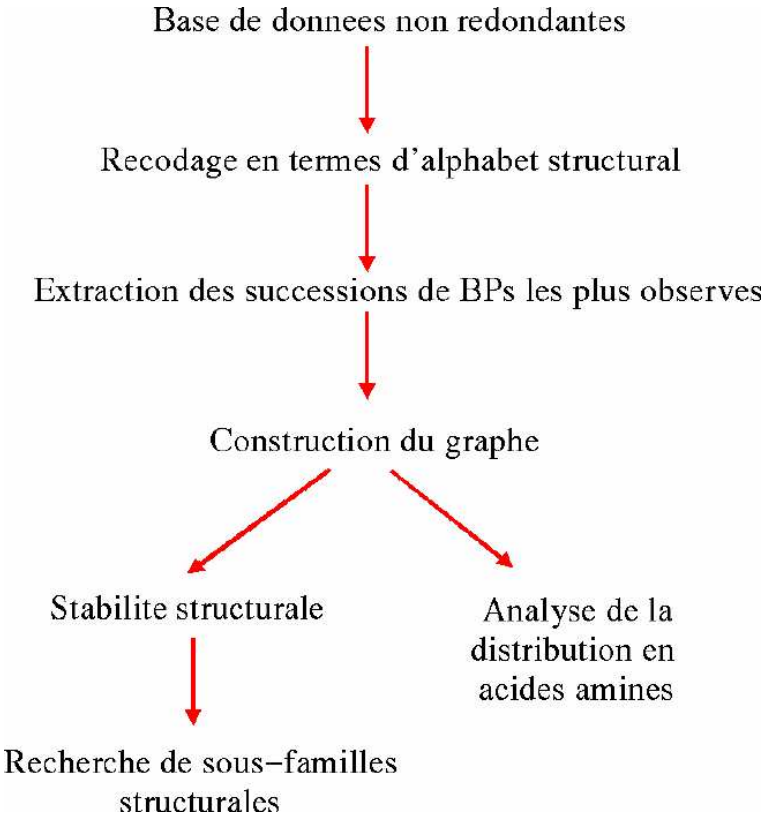


FIG. 5.6 – *Schéma récapitulatif de la validation du réseau discontinu.*

La prédiction de la structure tridimensionnelle d'une protéine à partir de sa séquence primaire est un défi constant de la biologie moderne. Pour apprécier l'hétérogénéité des structures protéiques, nous avons défini un ensemble de 16 prototypes (ou blocs protéiques) qui décrivent "au mieux" la structure locale des protéines [35]. En analysant une base de données de structures protéiques recodées à l'aide de ces blocs, certaines successions de blocs, i.e. des mots structuraux, apparaissent particulièrement fréquentes et fortement dépendantes. Elles permettent la construction d'un graphe orienté simple reliant de manière logique l'ensemble des motifs de 5 blocs les plus fréquents. Le réseau obtenu est composé de 31 nœuds distincts, correspondant

chacun à un des 16 blocs protéiques. Il contient du fait de sa conception, plus de 83% des résidus des protéines et point intéressant plus de 70% des boucles, régions considérées "variables". Pour évaluer la pertinence du réseau à reproduire des structures 3D homogènes, 17 motifs décrivant la totalité du réseau en ont été extraits. Ils ne correspondent qu'à un type unique de repliement. L'écart quadratique moyen observé *RMSd* est assez bon, il varie entre 1.37 à 3.07 Å, pour des tailles allant de 8 à 11 C_α. D'un point de vue angulaire, ces motifs sont aussi stables que les BPs seuls et ceci pour des tailles supérieures. L'utilisation de l'alphabet permet de se débarasser en fait des problémes liés à l'attribution des structures secondaires [31], principalement les feuillets β.

50 % de "structures secondaires répétitives se retrouvent en deux blocs. Le réseau pemet la construction des voies "logiques" qui entrent et sortent des BP m et du BP d qui vont suivre aussi longtemps que possible ces chemins "préférentiels". En fait, ce réseau ne montre pas que les boucles sont mieux déterminés que ce que l'ont pensent, mais qu'en fait les catégorisation actuelles basées sur l'existence d'une classification en 3-états (hélices α, feuillets β et boucles ou plutôt non-α/non-β) entraînent une classification biaisée. Des approches classiques se basent sur des structures de taille constante comprises entre deux structures répétitives, alors que la définition des bornes est loin d'être optimale, et, donc en conséquence des parties importantes de ces boucles semblables (sur le début ou sur la fin principalement) sont déclarées dans des groupes de longueur différentes. Ce réseau montre, en outre, des chemins préferentiels ,mais pas toujours continu. Les taux de départ/fin des fragments dans la réseau montrent clairement que les fragments compris dans le réseau commencent un peu partout et finissent un peu partout (cf. figure 5.2), conservant toutefois un fort déterminisme structural (cf. tableau 5.1). Les taux importants liés au nœud *c 05* présent dans la succession *mnopacd* (nœuds 01-02-03-05-06-07) sont dûs à la présence des autres blocs *c* (nœuds *12*, *22* et *23*) qui servent de terminateurs obligatoires. Cette notion de discontinuité doit être signalée pour ne pas prendre ce réseau comme un graphe continu.

Ainsi, on peut voir que l'alphabet même imparfait permet de retrouver des zones structuralement très stables et connnectées entre elles, dans les protéines. De plus, elles montrent localement des différences en répartition d'acides aminés ce qui est important pour la prédiction. L'analyse des fréquences des acides aminés associées à chaque partie du réseau permet de voir des différences séquentielles importantes. Ces résultats devraient être mis en relation avec la

méthode de prédiction bayésienne. Cette approche permettra une amélioration des recherches *ab initio* de la structure 3D à partir de la séquence. D'ailleurs, nous avons depuis développé une nouvelle méthode de prédiction basée sur l'utilisation de ces voies. Cette nouvelle stratégie nommée stratégie d'*épinglage* permet de prédire localement un mot structural et ensuite de l'étendre. Le taux de prédiction global passe de 40,7 à plus de 45 %. De plus, une amélioration de cette technique nommée *multi-épinglage* permet de générer plusieurs longs fragments en chaque position de la séquence. L'ensemble de ces deniers résultats permet de valider les concepts que nous avons développé.

Chapitre 6

Compactage des structures tridimensionnelles des protéines

6.1 Objectif

Dans ce chapitre, nous présentons une nouvelle approche de compactage des structures 3D des protéines, labellée méthode de la protéine hybride (MPH). L'objectif de MPH répond à un problème lié aux principes même des méthodes de classification. Dans une méthode de classification par partition type nuées dynamiques ou carte de Kohonen, le nombre de classes est fixé, et ces classes sont disjointes. La proximité entre les classes obtenues ne peut être vu qu'*a posteriori*. Les chaînes de Markov Cachées [149] sont une alternative à ce type de problème, toutefois, elles se basent sur la donnée des lois *a priori* des observations, ce qui est génant quand les lois appropriées ne sont pas connues. Appliqué à notre recherche, cela revient à considérer indépendant chacun des blocs, alors qu'une continuité existe entre un bloc *y* et un bloc *y'*. Le principal intérêt de MPH sera de mettre directement en relation les groupes pour tenir compte de la continuité. Le réseau discontinu (cf. chapitre 5) nous a démontré que des séries de blocs protéiques existent, avec MPH, nous allons observer des séries "floues" de blocs protéiques et ainsi analyser si des séries telles que, par exemple, *uvwxyz* et *uv'wxy'z* sont réelleement distinctes. Avoir un alphabet structural plus long (des séries de blocs protéiques) et plus "flou" (*uvwxyz==uv'wxy'z*), nous permetterait en effet d'avoir des séries équivalentes plus nombreuses, donc de réduire par la suite le nombre de fragments à utiliser pour des méthodes de recherche d'homologie structurale ou d'enfilage.

6.2 Principe général de la méthode de la protéine hybride

Dès mon DEA, avec le Pr. Hazout, nous avons travaillé sur cette problématique. Notre méthode se base sur une observation: "dans les protéines, s'il y a une entrée dans un feuillet, il y a une partie centrale puis une sortie qui va aller vers une boucle, et ainsi de suite jusqu'à la fin de la protéine". Il y a donc une continuité logique, une architecture locale forte. Ainsi, quand on fait une liste des séries de blocs, on observe des successions préférentielles, ceci permet même la construction d'un graphe, d'un réseau exprimant ces transitions (cf paragraphe 5.2).

En conséquence, il est possible de concevoir qu'une série X_i d'observations, peut aller vers une série X_{i+1}, et celle-ci ira vers une série X_{i+2}, et ainsi de suite. De même, la série X_i va aller vers une série X'_{i+1} qui ira vers X_{i+2} ensuite; X_{i+1} et X'_{i+1} se différenciant fort peu. Dans notre cas, par exemple, cela peut correspondre à des sorties de feuillet qui sont légèrement différentes. Avec une méthode classique de partitionnement, il est difficile de ne pas surdécouper le problème et ainsi considérer X_{i+1} et X'_{i+1} distincts.

Aussi, pour appréhender ce problème, nous avons élaboré une méthode dite "Méthode de la Protéine Hybride" (en francais, MPH, en anglais, "Hybrid Protein Model" ou HPM [36, 37]). La protéine hybride est une matrice de longueur L et de dimension l, cette dernière étant la dimension des observations (par exemple, 20 si l'on travaille sur les acides aminés ou 16 pour les blocs protéiques). En chaque position j, au départ, pour chaque type d'observation, une valeur moyenne est mise (la fréquence des acides aminés ou des blocs par exemple), ce qui représente une valeur attendue aléatoirement.

Le principe de l'apprentissage consiste à tirer des observations de longueur p aléatoirement dans la base de données, et, à les placer au mieux dans la matrice par une technique proche des cartes auto-organisées. Ainsi, petit à petit les tendances se focaliseront. Avec des paramètres d'apprentissage correctement choisis, à la fin de l'apprentissage, toutes les observations ressemblant à X_i seront en un site donné j_i et celles ressemblant à X_{i+1} et X'_{i+1} seront en j_{i+1}. En conséquence, il y a un regroupement local qui permet, si un apprentissage a été effectué avec des observations de longueurs p, d'avoir des successions qui permettent une analyse d'un ordre supérieur à p.

Cette approche a été appliquée aux séries de blocs protéiques (cf. paragraphe 6.3) et aussi lors

d'un apprentissage commun entre la séquence et la structure à l'aide de descripteurs physico-chimiques et d'angles dièdres (cf. parapgraphe 6.5). Cette approche a été appliquée à l'analyse de données génomiques par un Chromosome Hybride et a donné des résultats intéressants sur les régions subtélomériques des chromosomes de la levure, dont les réarrangements sont particulièrement étudiés [16, 1].

6.3 Application au compactage des structures

6.3.1 Objectif

Prédire la structure tridimensionnelle des protéines à partir de la séquence n'est pas une tâche aisée. Pour des protéines ayant des taux d'identité de séquences supérieurs à 30%, diverses techniques existent, elles se basent principalement sur des modélisations avec contraintes spatiales, physico-chimiques et des méthodes statistiques [15, 169, 93]. L'absence de protéines proches implique l'utilisation de méthodes différentes comme la prédiction *ab initio* (cf. paragraphe 2.2.2.3), qui toutefois est limitée à des protéines de petite taille [39, 137]. D'autres approches existent comme l'enfilage (ou "threading", cf. paragraphe 2.2.2.2) qui consiste en une recherche de compatibilité entre une séquence et un grand nombre de fragments issus de structures de protéines connues [126, 153, 7, 110]. Toutes ces techniques nécessitent l'utilisation d'une base de données protéiques non redondante (cf. paragraphe 2.2.1.4). Actuellement sont disponibles des bases de données reposant sur un critère de similitude de séquence [84, 83], et de similitude structurale [126, 87]. Pour cette première approche, MPH va donc servir à la compression d'une base de données structurales en une protéine hybride, ce qui pemettra dl'obtention d'un nombre limité de conformations locales, de repliements. Réduire le nombre de conformations locales peut être intéressant pour des recherches comme l'enfilage ou la recherche d'homologie structurale.

6.3.2 Principe

Ce chapitre va donc mettre en avant l'utilisation de MPH dans la compression d'une base de données structurales en une protéine hybride. Dans l'étape d'apprentissage, pour pouvoir compacter la structure, nous avons utilisé la traduction des protéines de cette base structurale

en suite de blocs protéiques. Chaque bloc représente toujours 5 C_α consécutifs, l'utilisation de séries de blocs va permettre la concaténation de différents repliements.

La compaction des structures générera des structures locales avec des parties communes et d'autres moins déterminées. Cette utilisation de séries de blocs dans la protéine hybride est à mettre en relation avec le fait que le choix de la longueur et le nombre des blocs a toujours été une question ardue donnant lieu à de nombreuses réponses : 6 blocs longs de 7 résidus pour Fetrow et collaborateurs [53], 4 à 7 blocs pour une longueur de 4 résidus pour Rooman et collaborateurs [162], 12 blocs ayant 4 résidus pour Camproux et collaborateurs [23, 24], 13 blocs pour des longueurs variables pour Bystroff et Baker [19], et une centaine d'hexamères pour Unger et collaborateurs [198] et Schuchhardt *et al.* et collaborateurs. Les structures locales comprises dans la protéine hybride nous permetterons de passer au-dessus de cet écueil en proposant des fragments longs, mais composés de blocs plus courts.

6.3.3 Application du Modèle de la Protéine Hybride aux séries de blocs protéiques

Pour compacter notre base de données structurales, nous avons donc utilisé la méthode de la protéine hybride (cf. paragraphe 6.2). La protéine hybride est une protéine chimérique composée de N sites où chaque position i n'est pas définie par un seul bloc, mais par une loi de probabilité $f_i(b_x)$, b_x étant un des 16 BPs (x=1, 2,..., 16).

La figure 6.1 récapitule les différentes étapes de l'apprentissage. Dans un premier temps, la base de données de 553 protéines ayant moins de 25% de similitude de séquence (cf. paragraphe 3.3.2.6) établie par Romain Gautier a été traduite en terme de blocs protéiques.

Ensuite, l'apprentissage proprement dit consiste en une recherche de la position optimale de chaque fragment dans la protéine hybride (cf. figure 6.1c à 6.1f). L'apprentissage est proche des cartes auto-organisées (cf. Annexe 1), mais sans processus de diffusion. Ainsi, la protéine hybride correspond à des séries de structures locales "floues".

L'intérêt principal de la méthode est de maintenir la séquentialité pour créer des séries de structures locales successives ayant un fort taux de recouvrement. Cette stratégie diffère donc fortement d'une classification classique où les groupes sont indépendants.

Un fragment $\boldsymbol{F}$ est tiré aléatoirement dans la base de données (cf. figure 6.1c). Chaque

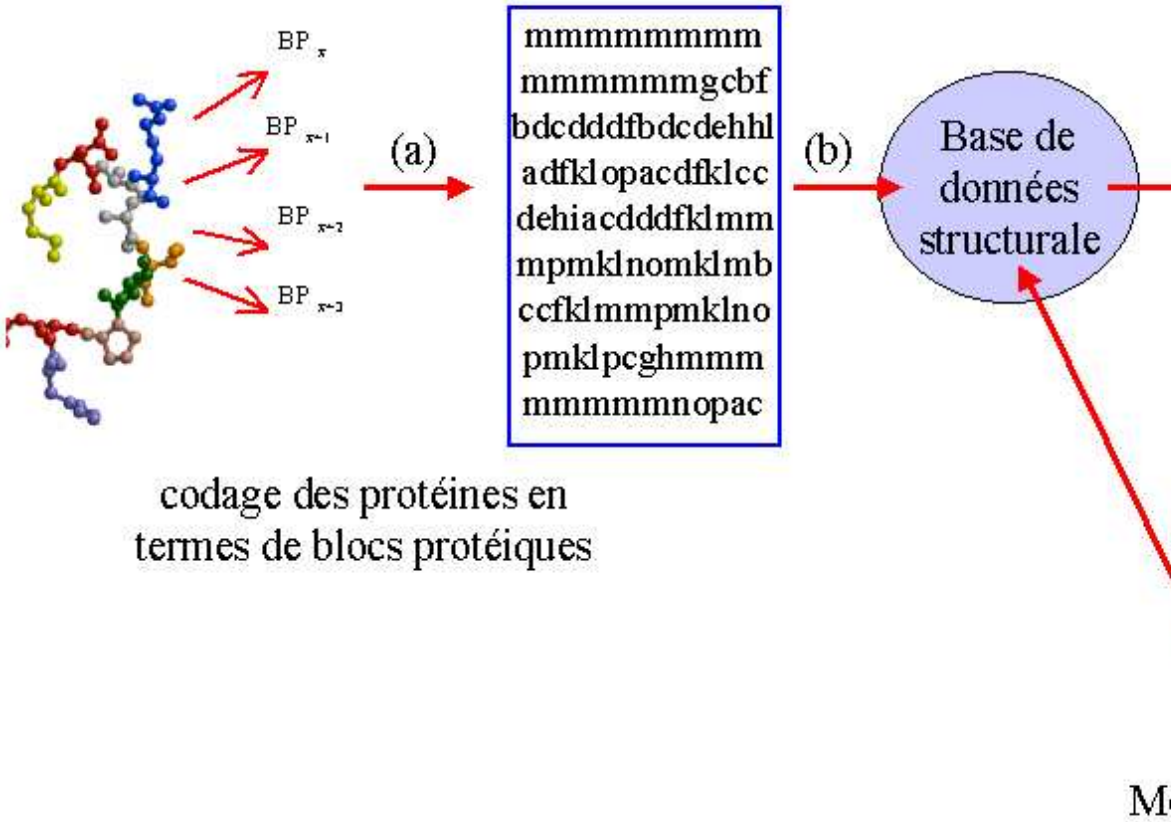


FIG. 6.1 – *Principe de l'apprentissage de la protéine hybride des blocs protéiques. en blocs protéiques . (b) Constitution d'une base de données structurales. (c) similitude en chaque position de la protéine. (e) Recherche de la position optima (c) jusqu'à la stabilisation du système.*

structure locale $\boldsymbol{F}$ est définie par L blocs consécutifs $\{b_1^*, b_2^*, ..., b_L^*\}$ (ici L=10 BPs, soit 14 C_α).

Localisation de la structure locale dans la protéine hybride. Un score S_i est calculé en chaque position i de la protéine hybride :

$$S_i = \sum_{k=1}^{k=L} ln[f_{i+k-1}(b_k^*)]$$

avec k=1,2,...,L. Ainsi le score S_i (i.e. le logarithme de la vraisemblance d'observation de la structure locale $\boldsymbol{F}$ en un site donné i) mesure la similitude entre la structure locale et une région donnée de même dimension dans la protéine hybride (cf. figure 6.1d). La région la plus proche de la structure locale possède le moins de différence avec celle-ci et donc son score est maximal. Ainsi cette position i_0 est définie comme (cf. figure 6.1e):

$$i_0 = \ argmax[S_i]$$

Modification locale de la protéine hybride. Ayant trouvé une zone de plus forte ressemblance, cette zone va être légèrement modifiée pour ressembler à la structure locale $\boldsymbol{F}$. Les positions i_0 à i_0+L-1 seront ainsi modifiées (cf. 6.1f):

si $b = b_k^*$ (i.e. le bloc protéique présent à la position k dans la structure locale), alors

$$f_{i_0+k-1}(b) \leftarrow \frac{f_{i_0+k-1}(b) + \alpha}{1+\alpha}$$

si $b \neq b_k^*$ (i.e. les autres blocs protéiques présents à la position k), alors

$$f_{i_0+k-1}(b) \leftarrow \frac{f_{i_0+k-1}(b)}{1+\alpha}$$

Le symbole $\leftarrow$ signifie que la valeur calculée remplace la valeur précédente. Le coefficient d'apprentissage α est encore égal à $\alpha_0/(1 + t/\nu)$, avec α_0 le taux initial d'apprentissage (ici $\alpha_0 = 0.1$), t le nombre de structures locales de L blocs déjà utilisées dans l'apprentissage et ν le nombre de structures locales présentes dans la base de données. L'apprentissage est progressif, ainsi l'ensemble de la base de données a été examiné entièrement C fois, ici C = 15 (phases: figure 6.1c à la figure 6.1f pour chaque structure locale de la base de données). Un passage complet de la base de données est appelé cycle. Pendant le premier cycle, le coefficient

d'apprentissage α est maintenu constant ($\alpha = \alpha_0$) pour modifier de façon importante la protéine hybride et ainsi faire moins jouer l'initialisation [109].

Initialisation. La protéine hybride est initialement définie par une série de N distributions sur les blocs $f_i(b_x)$ quasiment identiques :

$$f_i(b_x) = f(b_x).(1 + \epsilon_i)$$

avec $f(b_x)$ la fréquence du bloc protéique b_x dans la base de données, ϵ_i est une valeur aléatoire tirée dans l'intervalle [-τ;+τ] (τ a été fixé à 0,20). En chaque position la somme des probabilités des blocs a été réajustée à 1, en recalculant :

$$f_i(b_x) \leftarrow \frac{f_i(b_x)}{\sum_{i=1}^{i=16} f_i(b_x)}$$

Il faut noter que la protéine hybride est "fermée" dans le sens où le Nème site est contigu avec le premier, ainsi, il n'existe pas d'effet de bord.

6.3.4 Résultats

6.3.4.1 La protéine hybide

La figure 6.3 donne les résultats de l'apprentissage de la protéine hybride après $C = 15$ cycles à partir d'une protéine hybride initiale avec un coefficient d'initialisation $\tau = 0{,}20$ (cf . figure 6.2) et un coefficient d'apprentissage $\alpha_0 = 0{,}10$. La figure 6.3a montre la composition en blocs le long de la protéine hybride. L'analyse de la protéine hybride montre que les structures secondaires répétitives sont clairement détectables (les hélices α avec le bloc m et les feuillets β avec le bloc d). Trois types d'hélices α sont distinguables par leurs tailles : 2 à 4 BPs (positions [38:41]), 7 BPs [3:9] et 10 BPs [82:91], et 4 pour les feuillets β: 4 BPs [66:69], 5 BPs [74:78], 8 BPs [51:58], 9 BPs [15:23].

Différentes transitions entre structures secondaires régulières sont trouvées:

- hélice α à hélice α en positions [93:96],
- hélice α à feuillet β [9:14],

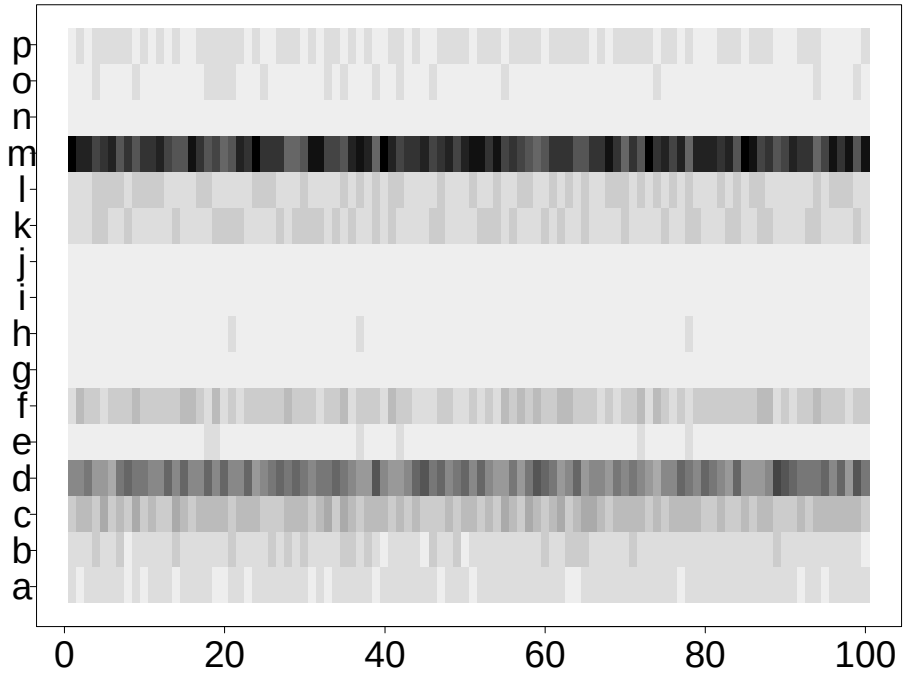


FIG. 6.2 – *Protéine hybride initiale. En abscisse les N sites, en ordonnée les 16 BPs, en chaque position se trouve la fréquence du BP correspondant avec une variation autorisée de τ = 0,20. L'ensemble des fréquences par site est renormalisé à 1,0.*

- feuillet β à hélice α [33:37] et [99:1],

- feuillet β à feuillet β [23:28], [58:65] et [69:71].

La spécificité des sites est nette, en chaque site seuls deux ou trois types de blocs sont présents. De plus, la majorité des structures locales sont consécutives. Ainsi, quand une structure locale $\boldsymbol{F}_j$ est localisée en position i_0 sur la protéine hybride, alors la structure locale $\boldsymbol{F}_{j+1}$, qui suit la structure locale $\boldsymbol{F}_j$, est localisée en la position i_0+1 sur la protéine hybride avec une probabilité de 81 %. La continuité est uniformément présente, aucune zone de cassure particulière n'est visible. Les structures locales sont réparties de manière assez uniforme autour d'une moyenne de 1115 observations par site (cf. Figure 6.3b), seules les positions des hélices α régulières sont largement au-dessus.

Il faut bien noter que chaque site i de la protéine hybride est un ensemble de structures locales de longueur L, ainsi les sites *i-1* et i ont *L-1* distributions de blocs en commun.

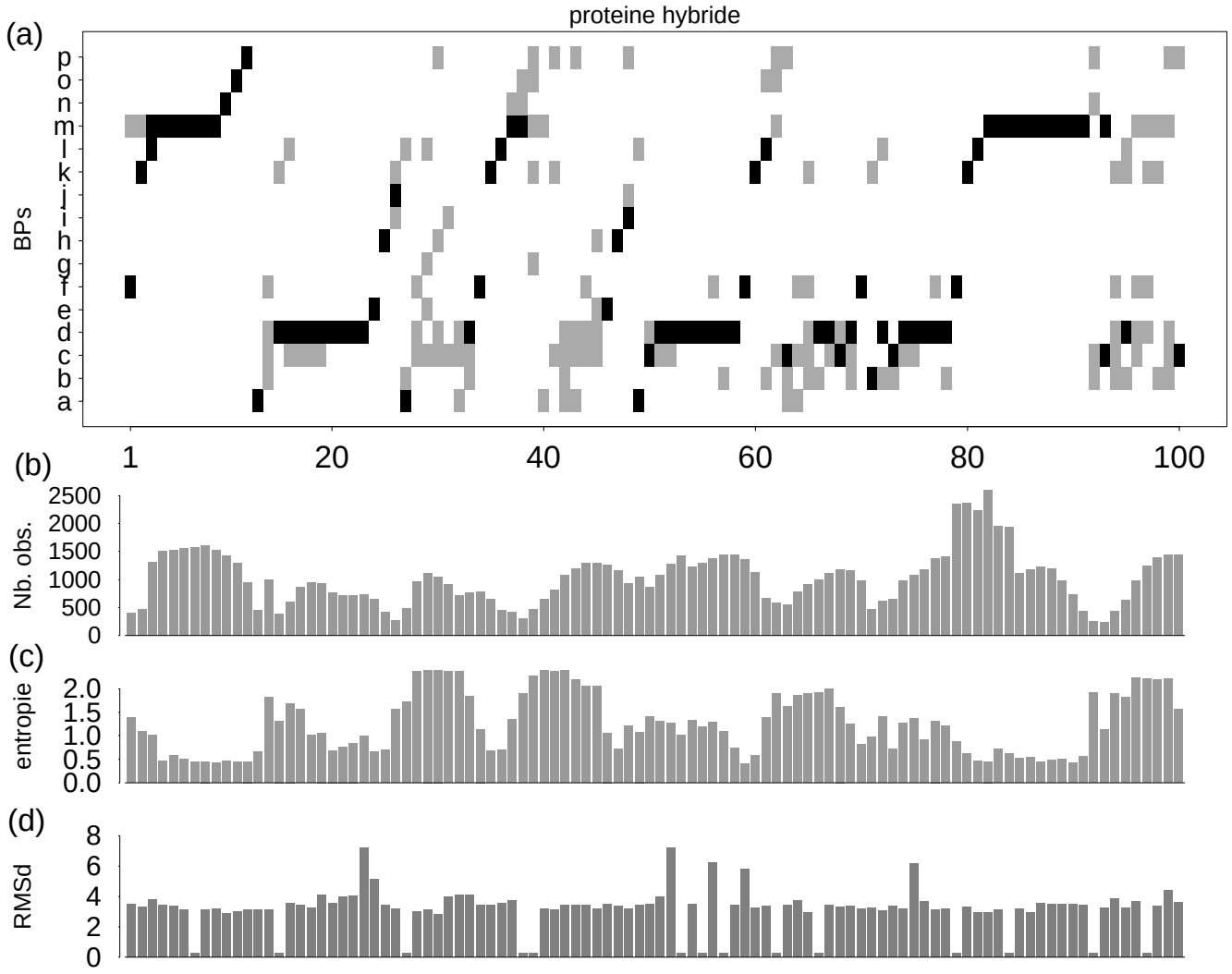


FIG. 6.3 – *Résultats de l'apprentissage. (a) La protéine hybride avec en chaque position, en noir, les fréquences des blocs supérieures à 0,35; en gris, ceux compris entre 0,35 et 0,10; en blanc, ceux inférieurs à 0,10. (b) Le nombre d'observations, (c) l'entropie relative et (d) le* RMSd *moyen de chaque site.*

6.3.4.2 Calcul de l'entropie des distributions des blocs protéiques

Chaque site de la protéine hybride est défini par une loi de probabilité des blocs. Une entropie H_i peut donc être calculée pour quantifier la diversité des blocs en chaque site :

$$H_i = -\sum_{b=1}^{16} f_i(b).ln[f_i(b)]$$

avec i la position du site, b un type de blocs protéiques et f_i sa distribution correspondante. Ainsi, une entropie faible correspond à un site avec peu de blocs hautement probables et donc hautement déterminés.

6.3.4.3 Analyse de l'entropie

L'entropie calculée le long de la protéine hybride est représentée dans la figure 6.3c et montre la haute spécificité de chaque site avec une valeur maximale de 2,40 et minimale de 0,41, 40% des positions ont une entropie inférieure à 1,0. Trois catégories de sites peuvent être distinguées par leur entropie.

Un premier groupe dont l'entropie est inférieure à 1,0. Il contient les sites les moins variables comme les hélices α. Ainsi, pour les positions [4:13] une hélice α centrale est suivie par une sortie C-terminale, les structures locales commencent principalement avec 6 BPm suivis par BPn, BPo, BPp et BPa aux sites [10:13], ceci est alors noté m_6nopa. Les structures locales en positions [79:91] sont des hélices α ayant une extrémité N-terminale $fklm_{10}$. Comme l'apprentissage s'effectue avec des structures locales de 10 blocs protéiques, ce motif contient en réalité les structures locales de type $fklm_7$, klm_8, lm_9 et m_{10}. La protéine hybride permet aussi la création de feuillet β avec extension N-terminale en [20:25] avec un motif d_4eh et des transitions courtes, telles des extrémités C-terminales de feuillet β dfk [58:60] et fk [35:36].

Le second groupe correspond à des zones intermédiaires avec une entropie comprise entre 1,0 et 1,5. Les plus longs correspondent à des feuillets β étendus ayant une extrémité N-terminale en positions [48:57], des structures locales de type iac_xd_{7-x},(x=1,2,3). Ensuite, les deux zones les plus représentatives correspondent à une extrémite N-terminale d'une hélice α en [1:3], fkl, et un feuillet β aux sites [77:78], d_2.

Le dernier groupe comprend les zones ayant une entropie élevée (supérieure à 1,5). Quatre zones sont plus variables et correspondent à des zones de boucles longues ou des feuillets allongés

en positions [26:34] et [38:45], des coudes entre deux feuillets [62:68] ou des boucles entre hélices α avec des feuillet β [94:100].

6.3.5 La stabilité structurale de la protéine hybride

La protéine hybride a donc bien appris l'ensemble des fragments, toutefois, ayant utilisé des structures locales de longueur égale à 10 blocs protéiques (la valeur de L), une question se pose quant à la qualité de l'approximation structurale de cet apprentissage.

Le *RMSd* moyen des fragments de chaque site a donc été calculé (cf. figure 6.3d). Le *RMSd* moyen est de 3,14 Å. Seuls 6 sites sont plus variables avec un *RMSd* moyen supérieur à 5 Å, et 14 sites ont un *RMSd* moyen inférieur à 1,0 Å.

Les zones structurellement variables sont principalement associées aux feuillets β. Par exemple, les structures locales en [19:28] correspondent à deux populations différentes: (i) un feuillet β court (type d_2) allant à un autre feuillet β, (ii) un feuillet β (de type d_4) allant vers une hélice α. De la même manière, les structures locales [48:57] et [52:61] sont associées avec des feuillets β de tailles différentes. Le site 59 correspond à des structures locales comprises entre [55:64] et composées de feuillet β de type $d_4fkopac$, $d_4fklpac$ ou c_2d_2fkopa. Le site 75 [71:80] est principalement composé d'une population de feuillet β $bccd_6f$ et d'une seconde population commençant par $bdcd_3$. En réalité, les zones les plus variables sont des feuillets β avec extrémité N- et C-terminale. Ces structures locales sont associées avec des séries de blocs plus complexes et diversifiées. Elles possèdent localement des séries communes et des parties distinctes, ce qui crée des zones particulièrement floues qui entraînent un *RMSd* moyen plus élevé.

Les structures locales ayant une variabilité faible sont des hélices α (voir tableau 6.1, 5 premières lignes) et quelques structures β (voir tableau 6.1, 5 dernières lignes).

En conclusion, les sites les mieux définis sont associés avec les structures régulières, mais pas exclusivement, comme avec le site 66 qui est une transition entre deux feuillets β.

6.3.5.1 Exemples de sites caractéristiques

Méthode d'analyse Pour quantifier l'importance de chaque type d'acides aminés en chaque position, les occurrences des acides aminés ont été calculées, en comptant pour chaque position

site	motif	caractérisation
7	$m_8$no	hélice α avec N- et C-ter
38	$fklm_8$	courte hélice α
79	d_4fklm_3	transition rapide entre un feuillet β et une hélice α
84	klm_8	une hélice α avec N-ter
92	m_5cmcd_2	hélice α allant à un feuillet β
15	opa	β N-ter
27	$ehia$	entrée en N-cap β
53	$acddd$	feuillet β C-ter
55	$acddd$—-f	feuillet β C-ter
57	$dddd$–f	feuillet β C-ter
62	opa	transition entre deux feuillets β
66	opa	transition entre deux feuillets β
97	cd —- f	feuillet β

TAB. 6.1 – *Sites de la protéine hybride les plus stables suivant le critère du* RMSd *moyen, avec la position des sites (centré sur le cinquième C_α), le motif associé localement et une description en termes de structures secondaires.*

s chaque série de L blocs protéiques. Elles sont ensuite normalisées en Z-scores :

$$Z_j^i = \frac{\left(n_j^i - n_a^i\right)}{\sqrt{n_a^i}}$$

où n_j^i est le nombre observé de résidus a en position j et n_a^i son nombre attendu. Ce dernier est le produit :

$$n_a^i = N_i.\mathbf{q}_a$$

avec N_i et $\mathbf{q}_a$ respectivement le nombre d'observations à cette position et la fréquence du résidu a dans la base de données. Ainsi, les Z-scores positifs (inversement négatifs) correspondent à des sur-représentations (inversement sous-représentations) de ce résidu en cette position (cf. paragraphe 3.3.5.2).

Exemples La figure 6.4 montre pour trois sites d'intérêt la superposition de fragments de longueur 10 C_α appartenant à chaque site et les matrices d'occurrence associées; ces matrices ont été normalisées en Z-scores. Ces sites sont dans l'ordre:

- le site 7 qui correspond aux structures locales [3:12], il a un *RMSd* moyen de 0,3 Å associé à une entropie de 0,43 (cf. figure 6.4a).

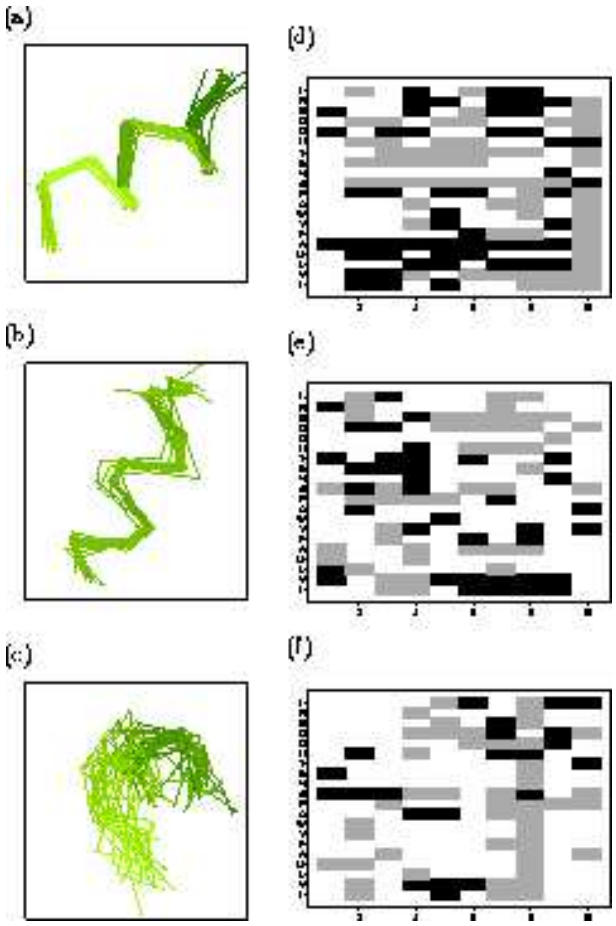


FIG. 6.4 – *Les sites caractéristiques en positions 7, 73 et 56, avec (a-c) les superpositions des fragments associés aux sites, visualisés par XmMol [194] et (d-e) les matrices d'occurrences associées à ces sites, normalisées en Z-scores (noir : Z-scores < 4,4, gris: Z-scores > -4,4 et blanc: Z-scores intermédiaires). Les acides aminés sont classés de bas en haut comme suit: I, V, L, M, A, F, Y, W, C, P, G, H, S, T, N, Q, D, E, R, K.*

- le site 73 [69:78] qui a une entropie de 0,72 et un *RMSd* moyen de 3,3 Å (cf. figure 6.4b).

- le site 56 [52:61], entropie de 1,39 et un *RMSd* moyen de 6,21 Å (cf. figure 6.4c).

Ces trois sites correspondent aux trois principales catégories de sites. Les figures 6.4d à 6.4f montrent les matrices d'occurrences associées normalisées en Z-scores. Le premier site (figures 6.4a et 6.4d) représente une partie centrale d'hélice α avec des sur-représentations en Alanine et des résidus non-polaires. La présence des résidus chargés tels la Lysine, l'Arginine et l'Acide Glutamique en positions 7 et 8 est classiquement associée avec la présence de l'extrémité C-terminale [154], et celles de Glycine et d'Asparagine en position 10 avec un élément de rupture structural. Le second site (figures 6.4b et6.4e) est quant à lui associé avec la présence d'acides aminés aliphatiques caractéristiques des feuillets β. L'extrémité C-terminale des feuillets β est vue avec des sous-représentations de résidus polaires en position 4 et en Isoleucine et Valine aux sites 5 à 9. Le dernier motif (figures 6.4c et6.4f) est moins déterminé, toutefois le repliement des structures locales est globalement similaire. En parallèle du manque de spécificité structurale, la composition en acides aminés est moins informative que précédemment. Seule la position 8 est hautement spécifique avec une sur-représentation de Glycine et d'Asparagine et une sous-représentation de 17 autres acides aminés.

6.3.6 Relation avec la séquence

6.3.7 Analyse de la répartition des acides aminés par l'utilisation des Z-scores

Ayant examiné précédemment trois sites caractéristiques, nous allons dans cette partie analyser la distribution des acides aminés sur l'ensemble de la protéine hybride pour caractériser la pertinence des sites en termes d'acides aminés et ensuite pour analyser leur relation avec les structures locales.

La figure 6.5 montre les acides aminés normalisés en Z-scores de la protéine hybride. Pour obtenir cette information, il suffit de compter en chaque position centrale de chaque fragment de L blocs protéiques le nombre d'occurrences de chaque acide aminé. Les proportions classiques des acides aminés des structures secondaires répétitives sont retrouvées avec les sur-représentations d'Alanine (positions [2:10] et [81:90]) pour les hélices α, des résidus chargés (Lysine, Arginine

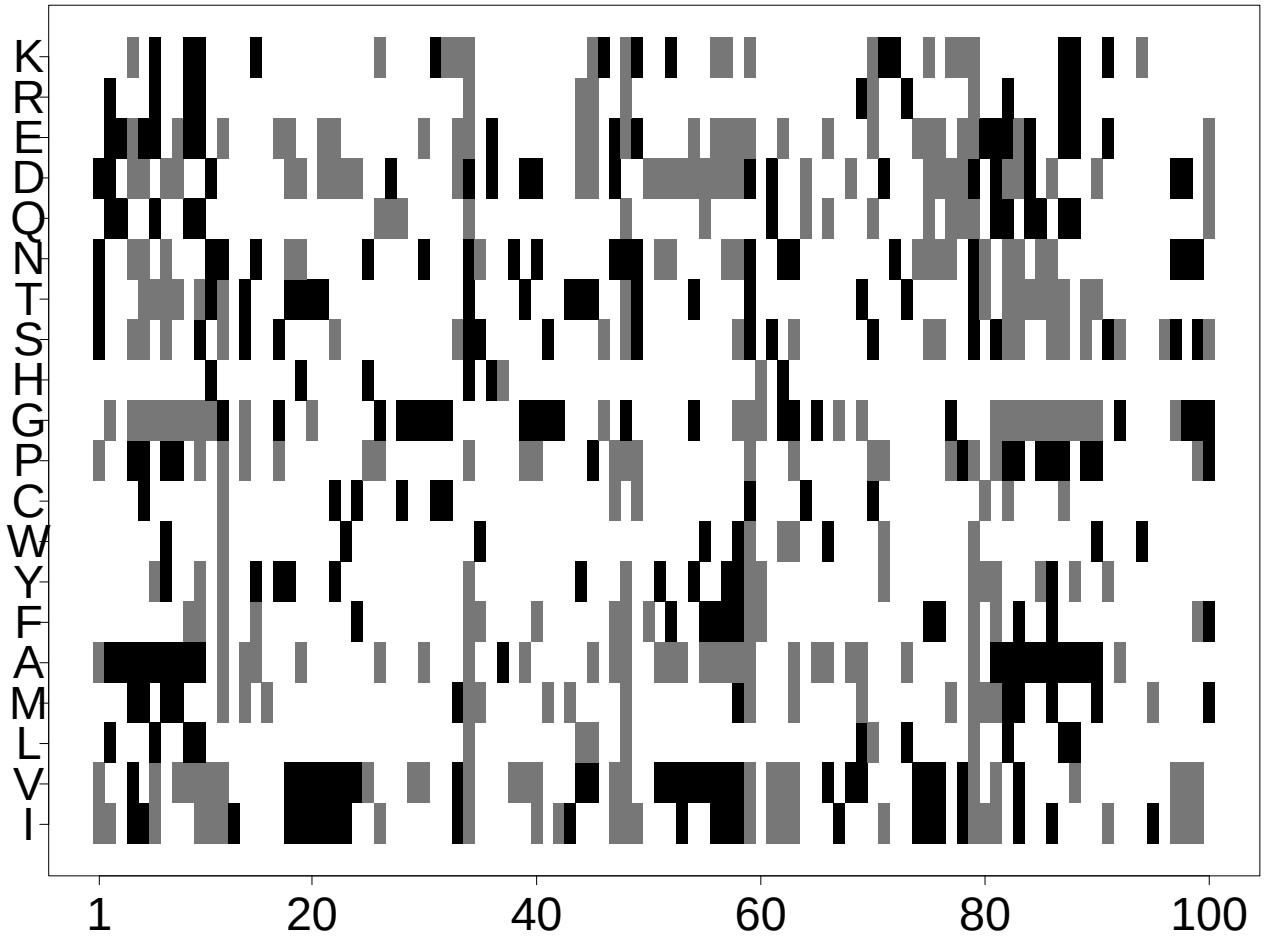


FIG. 6.5 – *Matrice d'occurrences des acides aminés centraux, normalisés en Z-scores (noir: Z-scores < 4,4, gris: Z-scores > -4,4 et blanc intermédiaire), avec en abcisse les 100 sites de la protéine hybride, en ordonnée les acides aminés.*

et Acide Glutamique) pour leur extrémité N-terminale (positions [9:10] et [87:88]), des résidus aliphatiques pour les feuillets β (positions [18:24] et [51:58]) et des Glycines (positions [28:32] et [39:42]) dans les boucles. Quelques spécificités intéressantes peuvent être observées comme la présence de Phénylalanine en extrémité C-terminale des feuillet β en [55:58], mais absente des autres extrémités C-terminales des feuillets β ([21:23], [69:70] et [76:78]).

6.3.7.1 KLd pour quantifier la spécificité des acide aminés

Les répartions des acides aminés attendues ont été retrouvées. Toutefois, elles ne donnent pas d'idées quant à la spécificité de chaque site d'un point de vue statistique. Aussi pour l'évaluer, les données précédentes ont été analysées á l'aide du KLd. Le calcul de l'entropie relative ou mesure de la divergence assymétrique de Kullback-Leibler (noté KLd [112]) a été explicitée dans le paragraphe 3.3.5.1. La spécificité des résidus a été calculée par rapport à ceux de la base de données. La quantité $N_i.K(\mathbf{p}_i,\mathbf{q})$ suit une distribtion de type χ^2 , avec N_i le nombre d'observations (ici, le nombre de structures locales associées au site i). Ainsi, les positions ayant une forte spécificité seront associées à des valeurs élevées.

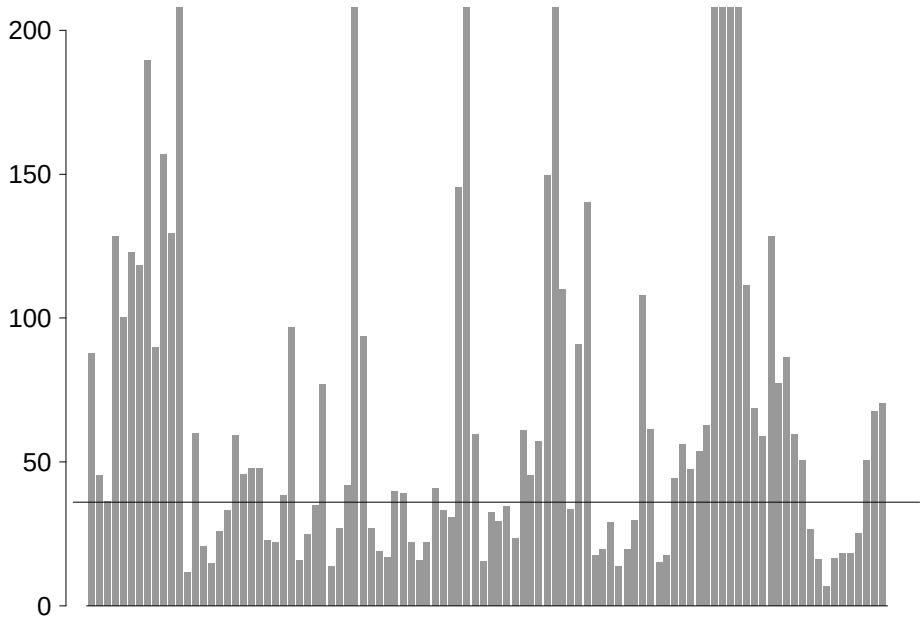


FIG. 6.6 – *KLd associé à la répartition en acides aminés (cf. figure 6.5) en chaque position de la protéine hybride, avec en abcisse les sites de la protéine hybride et en ordonnée la valeur de $N_i \times$ KLd, la valeur seuil est de 36 et représente une erreur de premier ordre de 0,05 avec 19 ddl.*

Un premier résultat visuel net apparaît: les valeurs obtenues sont nettement supérieures à celles calculées précédemment avec les blocs protéiques seuls (cf. Figure 6.6). La valeur seuil utilisée est de 36 soit un risque de première espèce (α) de 0,05 pour 19 degrés de liberté (les 20 types d'acides aminés). Les hélices α et les feuillets β sont présents, mais d'autres positions en dehors des structures répétitives présentent une structure stable comme des boucles en [59:60] incluant la série de blocs *kl*.

6.3.7.2 Similitude des sites de la protéine hybride en termes d'acides aminés

L'ensemble des sites ayant été précédemment caractérisé du point de vue de leur composition en acides aminés, se pose la question de la similitude existante entre ces sites. Pour voir si des sites ayant le même type de distribution du point de vue des acides aminés sont associés aux mêmes structures locales ou non, les sites ont été regroupés avec la méthode de classification *k-means* [79]. Cette méthode permet de regrouper dans un même groupe des observations proches (cf. Annexe 1).

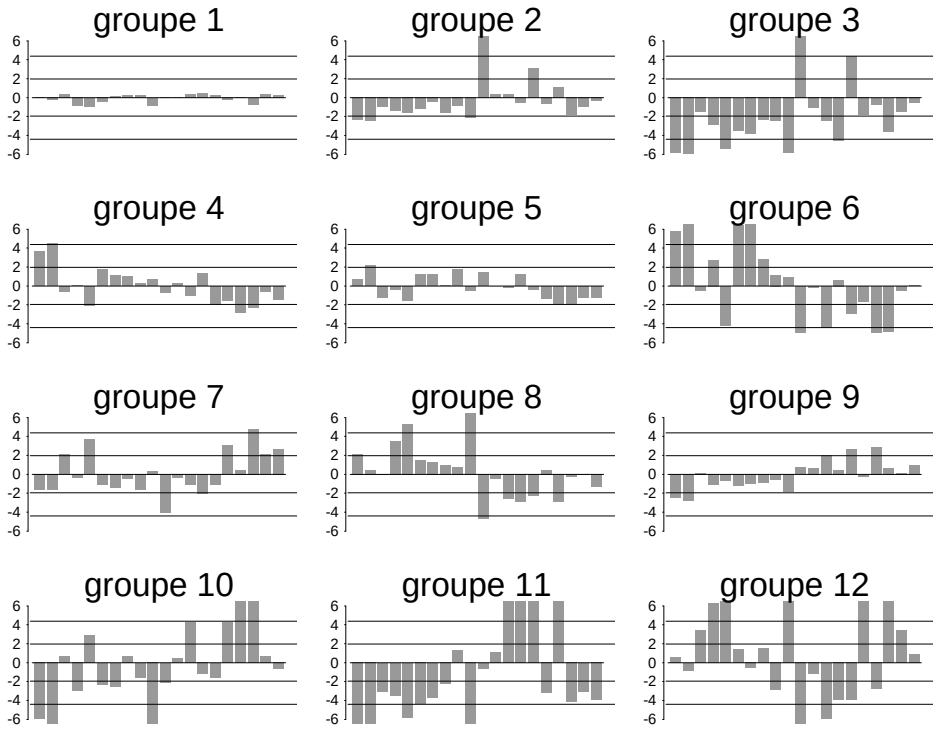


FIG. 6.7 – *Classification par la méthode des k-means des positions des acides aminés transformés en Z-scores (cf. figure 6.6). Les acides aminés sont classés de gauche à droite comme suit: I, V, L, M, A, F, Y, W, C, P, G, H, S, T, N, Q, D, E, R, K.*

Les 100 sites mis en Z-scores (cf. paragraphe 6.3.7) ont été classés en 12 groupes (logiciel R,

librairie *mva*, méthode *kmeans*) . La figure 6.7 présente les 12 groupes obtenus avec les valeurs pour ces 100 sites.

Le groupe 1 (25 sites) ne montre aucune spécificité en acides aminés et possède un nombre très divers de blocs pour chaque site, les 16 types de blocs protéiques se retrouvent associés à ce groupe. Le groupe 2 (5 sites) montre une forte sur-représentation en Glycine et Asparagine, cependant des blocs fort distincts lui sont associés comme les blocs d, j ou m. Le groupe 3 possède le même type de sur-représentation. Cependant il est fortement éloigné du précédent, car il est associé avec des sous-représentations plus fortes et ne correspond, par ailleurs, qu'à des blocs i, j et k qui correspondent à des zones de ruptures de structures régulières.

Les groupes 4 à 6 sont associés à des feuillets β, et les groupes 7 et 8 à des hélices α. Le groupe 4 (17 sites) est associé à une sur-représentation d'acides aminés aliphatiques tels l'Isoleucine et la Valine correspondant à des feuillets β centraux et des extrémités N-terminales de feuillet β. Le groupe 5 (12 sites) est composé de PB d et d'autres blocs associés aux extrémités N- et C-terminales des feuillets β (de PB a au PB f) et lié à une sur-représentation de Valine et une sous-représentation de résidus chargés. Le groupe 6 (1 site en position 58) est un feuillet β régulier avec une forte sur-représentation d'acides aminés non-polaires et une sous-représentation d'acides aminés polaires. Le groupe 7 (10 sites) correspond à une hélice α, plus spécifiquement, une extrémité C-terminale d'hélice α. Une sous-représentation de Leucine, Méthionine et de résidus polaires est retrouvée. Le groupe 8 (8 sites) montre une sur-représentation d'Alanine, de Méthionine, d'Isoleucine et de Proline et une sous-représentation de Glycine; ce groupe ne comprend que les parties centrales des hélices α.

Le groupe 9 (15 sites) avec une sur-représentation de petits acides aminés polaires est caractéristique des changements brusques de structures avec des PBs comme les PBs f, h, a, l et o. Le groupe 10 (1 site en position 81) est associé au PB h; il a une sur-représentation de résidus aliphatiques et de Proline, ainsi qu'une forte sous-représentation des résidus polaires. Le groupe 11 (3 sites) montre une sur-représentation de petits résidus polaires et des sous-représentation de résidus non-polaires et de Proline. Le groupe 12 (1 site en position 70) montre une sur-représentation de Glycine et une sous-représentation des résidus non-polaires. Pour les quatres sites qui composent les deux derniers groupes le bloc protéique le plus important est toujours le PBf, avec cependant une différence notable; dans le groupe 11, le bloc f est dans une série *fkl* alors que dans le groupe 12, il est dans une succession *fbd*. Ce détail est d'importance car,

comme l'a montré le graphe (cf. paragraphe 5.2), le premier fait partie d'une série allant vers des hélices alors que le second est plus présent dans les transitions vers des feuillets β.

6.3.8 Etude de l'influence des paramètres de l'apprentissage

L'influence des différents paramètres utilisés lors de l'apprentissage a été analysée :

(i) la longueur de la protéine hybride $N = 100$ a été choisie pour obtenir une caractérisation correcte des structures locales. Avec $N > 100$, certains sites auraient été fort peu peuplés, et pour $N < 100$, le nombre de structures locales mal approximées aurait augmenté fortement.

(ii) le coefficient d'apprentissage α_0 contrôle à la fois la qualité et la vitesse de l'apprentissage. α_0 pris égal à 0,10 permet de diminuer l'importance de l'initialisation, le système étant bousculé fortement pendant les premiers cycles d'apprentissage. L'utilisation d'une valeur plus faible et plus classique permet un apprentissage plus rapide mais moins sûr.

(iii) Le résultat de la protéine hybride n'est pas fortement modifié par le tirage des fluctuations aléatoires ϵ_i et du coefficient τ. Un léger décalage de la protéine hybride est souvent observé.

(iv) La valeur du nombre de cycles C est définie par l'utilisateur. En pratique, ce nombre est déterminé en se basant sur le fait que plus aucun changement notable n'est observé pendant l'apprentissage au delà d'un certain nombre de cycles. Dans la pratique, les 5 premiers cycles déterminent l'apprentissage dans les grandes lignes.

En conclusion, l'influence des paramètres est relativement mineur. La séquentialité qu'implique les blocs protéiques et la présence de structures comme les structures répétives et leurs entrées et sorties permettent d'obtenir des résultats stables.

6.3.9 Conclusion de l'apprentissage

MPH est donc une nouvelle approche qui permet le compactage d'une base de données de structures de protéines. L'étape d'apprentissage permet de conserver une excellente séquentialité entre les fragments (81% sont contigus). L'entropie montre que la caractérisation des

sites est correcte avec un maximum de trois types de blocs protéiques différents par sites. Le déterminisme étant important, des structures similaires sont classées dans les mêmes régions de la protéine hybride.

Le calcul des *RMSd* moyens a montré la forte stabilité des structures locales et ceci malgré la grande hétérogénéité de ces structures. Les *RMSd*s les plus importants sont dus principalement à des sites contenant à la fois des feuillets β courts et d'autres longs. Un point important est, que malgré un niveau élevé de *RMSd*, les repliements associés à ces sites possèdent des formes similaires. En outre, les structures répétitives ne sont pas les seules qui soient bien approximées.

L'exemple donné à la figure 6.5 montre bien trois catégories caractéristiques de sites avec leurs compositions plus ou moins marquées en acides aminés. Cette caractérisation en acides aminés montrée dans la figure 6.7 et la classification réalisée à partir de ces données montrent que plus de 75 % des sites sont informatifs. Ceci corrobore le fait qu'une partie des séquences protéiques ne sont pas informatives [177]. De plus, ces distributions en acides aminés sont souvent associées avec des distributions caractéristiques en blocs protéiques, ce qui peut être pris en compte dans le cas d'une prédiction de la structure par la séquence.

En conclusion de cette première partie, nous avons vu que la méthode de compaction par MPH permet d'obtenir une protéine hybride qui représente bien la majorité des structures locales. De plus, la composition en acides aminés est fortement spécifique. L'ensemble de ces données en prenant en compte le fait que la continuité est assurée fait de cette protéine un outil intéressant pour l'avenir. Dans les paragraphes suivants, la protéine hybride va servir à mettre en place une méthode rapide de recherche d'homologie structurale entre deux protéines.

6.4 Application à la recherche d'homologie

6.4.1 Principe de la recherche

L'intérêt principal de MPH appliqué séries de blocs protéiques est bien sur la compaction d'une base de donnée structurale. Toutefois, l'utilisation de la protéine hybride n'est pas limitée qu'à ce seul point. En chaque position, un ensemble de fragments proches se trouve localisé. Donc, si l'on recherche des fragments similaires courts d'un point de vue structural il suffit de trouver la position correspondante sur la protéine hybride. Pour trouver des fragments

plus longs, le principe est identique. Des fragments longs et similaires structuralement seront théoriquement localisés aux mêmes positions sur la protéine hybride. En outre, la continuité étant forte, ces positions seront souvent consécutives.

Nous avons donc mis au point une méthode qui tient compte de ces concepts pour rechercher des fragments structuraux proches. La figure 6.8 explicite l'ensemble du processus. Une première étape consiste à recoder la structure tridimensionnelle des protéines en blocs protéiques. A ce niveau simple, la recherche d'homologie est difficile, les structures répétitives gênant particulièrement la recherche. Aussi, la méthode utilise directement la protéine hybride.

Les séries de blocs, structures locales sont recodées en position sur le protéine hybride. Ensuite, un dotplot est calculé. Un dotplot est une matrice de taille N_1 par N_2 avec N_1 longueur de la première protéine sur N_2 longueur de la seconde protéine. Cette matrice est remplie comme suit :

si les positions sont les mêmes sur la protéine hybride $\rightarrow 1$.

si les positions sont différentes sur la protéine hybride $\rightarrow 0$.

Dans un premier temps, le dotplot est une matrice composée de 0 et de 1. Cette information doit donc être travaillée pour ne conserver que les plus longues successions identiques. Ainsi, le dotplot est filtré en ne conservant que les diagonales de longueur G où un nombre H de points est égal à 1. Travaillant sur des protéines proches, H a été pris égal à G. Au départ G a été pris élevé pour extraire les diagonales les plus longues, puis progressivement a été réduit. Les deux protéines étant proches et ayant des longueurs comparables entre 400 et 450 acides aminés, l'extraction des diagonales n'a été effectuée que dans une zone médiane [$-\theta$,$+\theta$]. Cela veut dire que pour une position c d'une protéine, la recherche de similitude s'est focalisée dans une région [c-θ,c+θ] de la seconde protéine et inversement. θ était fixé à 50 résidus. En outre, quand une diagonale était sélectionnée, les zones correspondantes dans les deux protéines n'étaient plus réutilisées ce qui permet de construire des véritables séries de fragments similaires distincts.

Enfin, pour vérifier que le fait de passer par les blocs protéiques, puis le recodage par la protéine hybride n'entraîne pas de faux positifs, chaque couple de fragments similaires a été superposée et le *RMSd* résultant calculé. Cette approche permet ainsi la recherche de structures possédant un repliement proche.

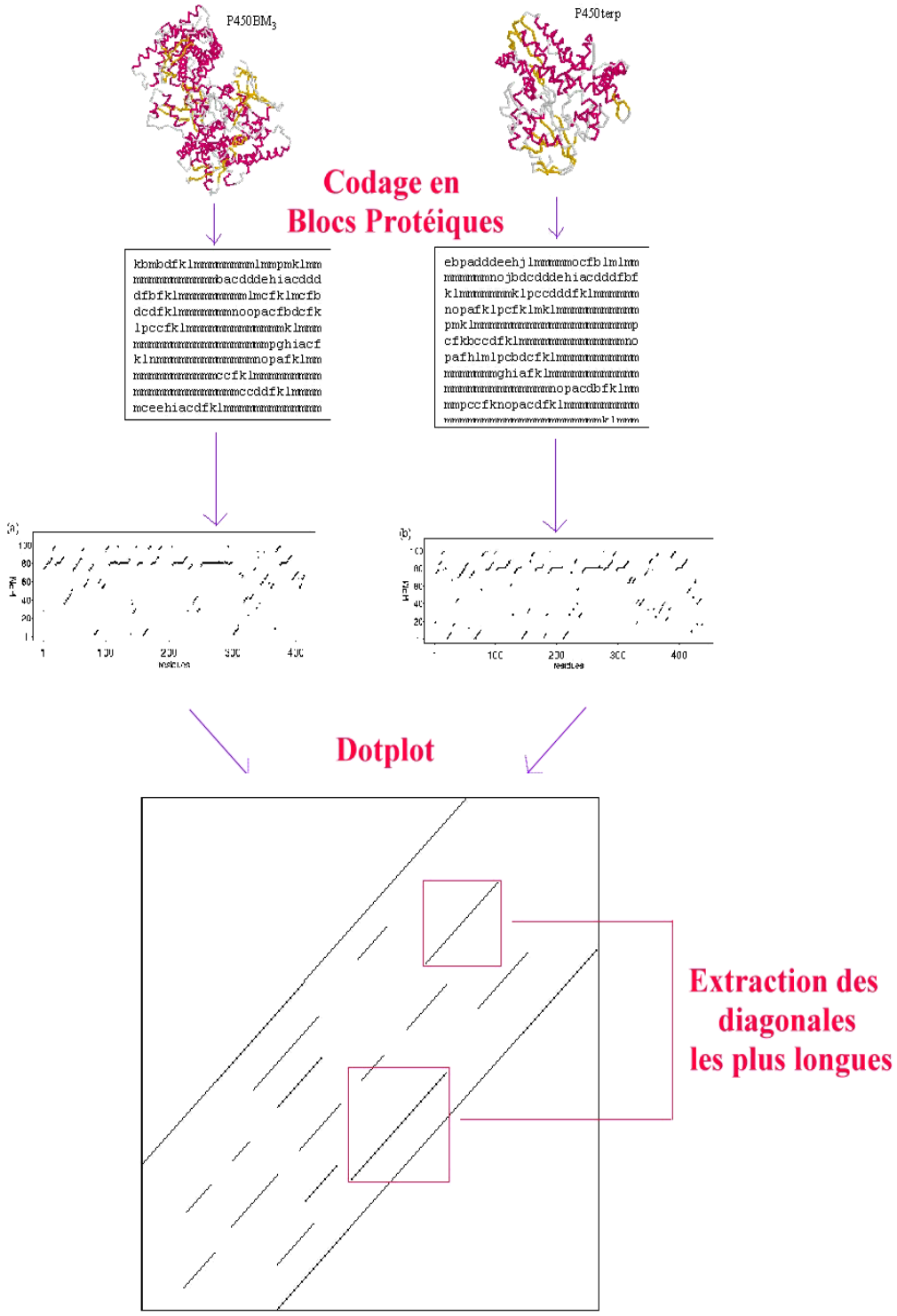


FIG. 6.8 – *Principe de la recherche d'homologues structuraux. Avec de bas en haut, le codage des deux structures tridimensionnelles en blocs protéiques, puis la localisation des séries de blocs protéiques sur la protéine hybride et enfin le dotplot des positions sur la protéine hybride des deux protéines et enfin la recherche des fragments les plus longs en commun.*

6.4.2 Homologie structurale pour deux cytochromes P450

Les cytochromes P450 ont été bien décrits. Ils ont une identité de séquence allant de 10 à 30%. Haseman et collaborateurs [80] les ont caractérisés en termes de fragments communs de structures secondaires (hélices α, 3_{10}-helix, π-helix, feuillet β et β-bugle). Jean et collaborateurs [94] ont déterminé des blocs structuraux communs (Common Structural blocs ou CSBs) entre differents cytochromes P450. Ils ont utilisés des comparaisons deux à deux pour ainsi aboutir à la modélisation d'un nouveau cytochrome P450. Nous avons utilisé deux cytochromes P450: $P450_{BM3}$ (code PDB: 2hpd[151]) et $P450_{terp}$ (code PDB: 1cpt[81]).

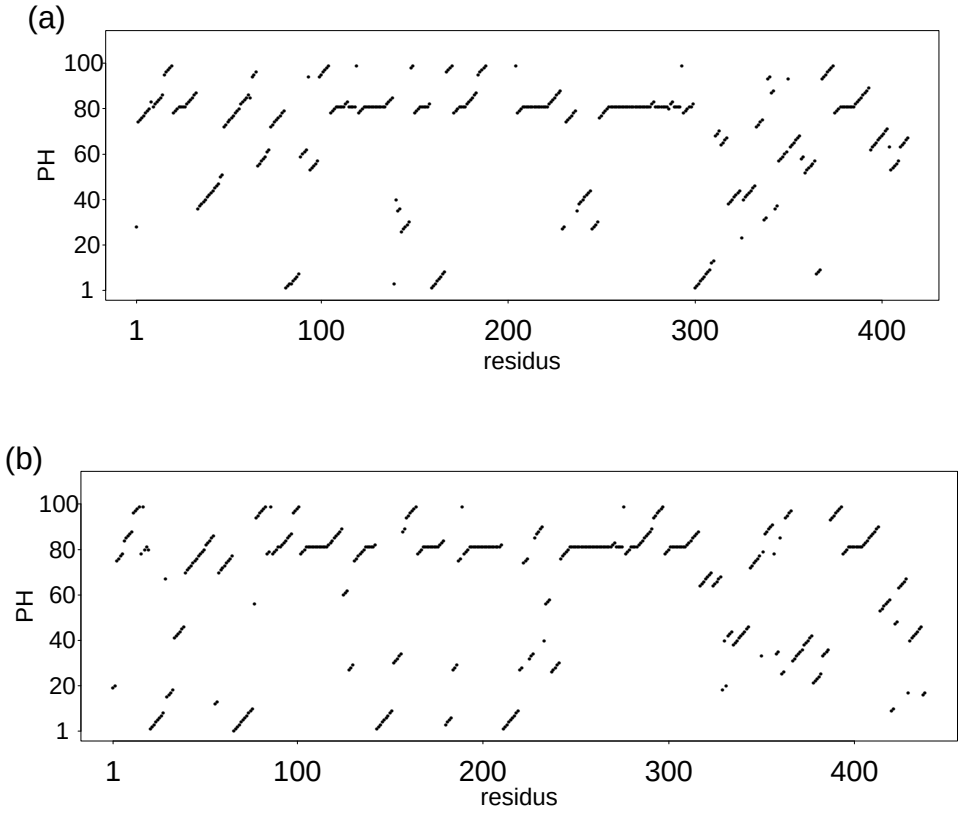


FIG. 6.9 – *Recodage (a) du cytochromes $P450_{BM3}$ et (b) du cytochromes $P450_{terp}$ sur la protéine hybride. Avec en abcisse les positions des résidus des cytochromes et en ordonnée de la protéine hybride.*

Dans un premier temps, les deux structures ont été codées en termes de blocs protéiques. Ensuite, utilisant des successions de 10 blocs, chaque protéine a été codée à partir de la protéine

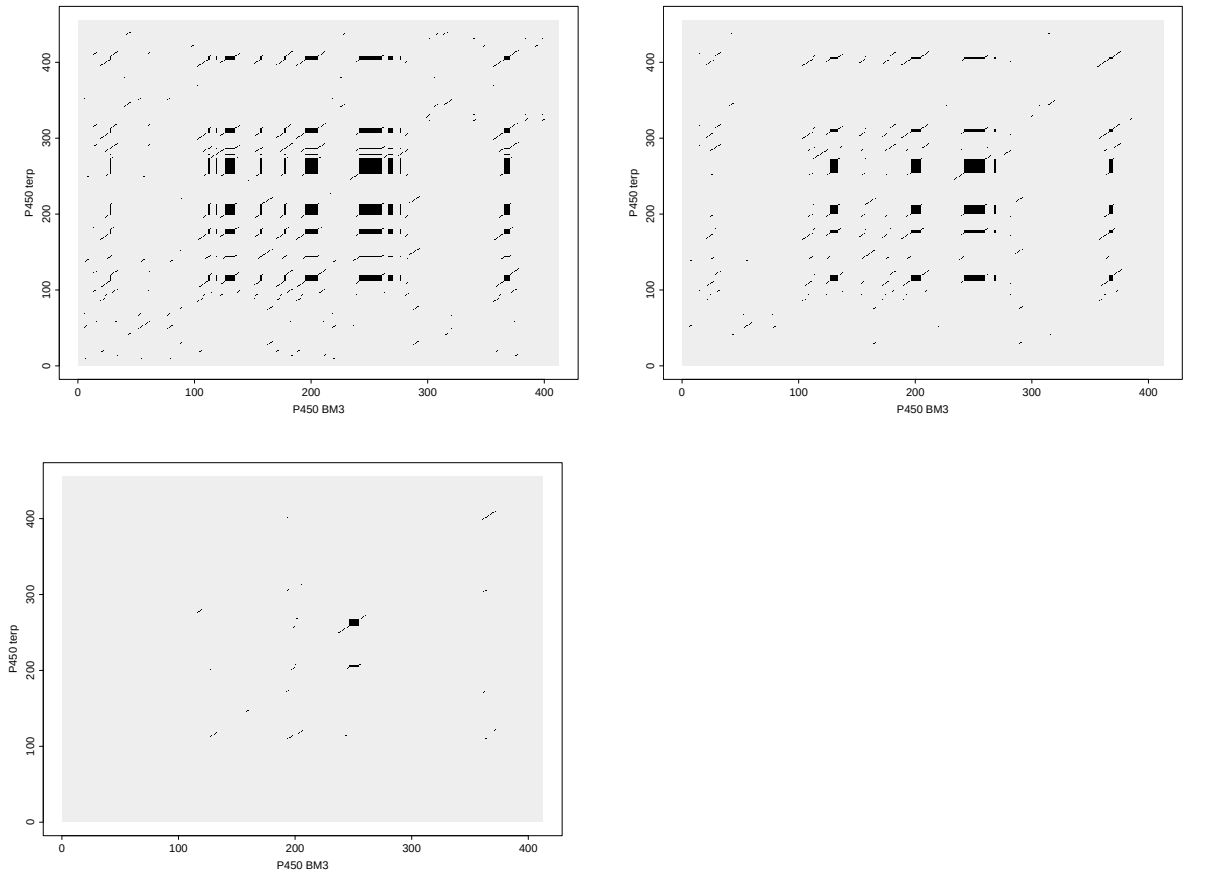


FIG. 6.10 – *Dotplot entre les positions recodées sur la protéine hybride entre $P450_{BM3}$ (en abcisse) et $P450_{terp}$ (en ordonnée) (cf. figure 6.9) avec un filtre (a) de taille 4, (b), de taille 6 et (c) de taille 15.*

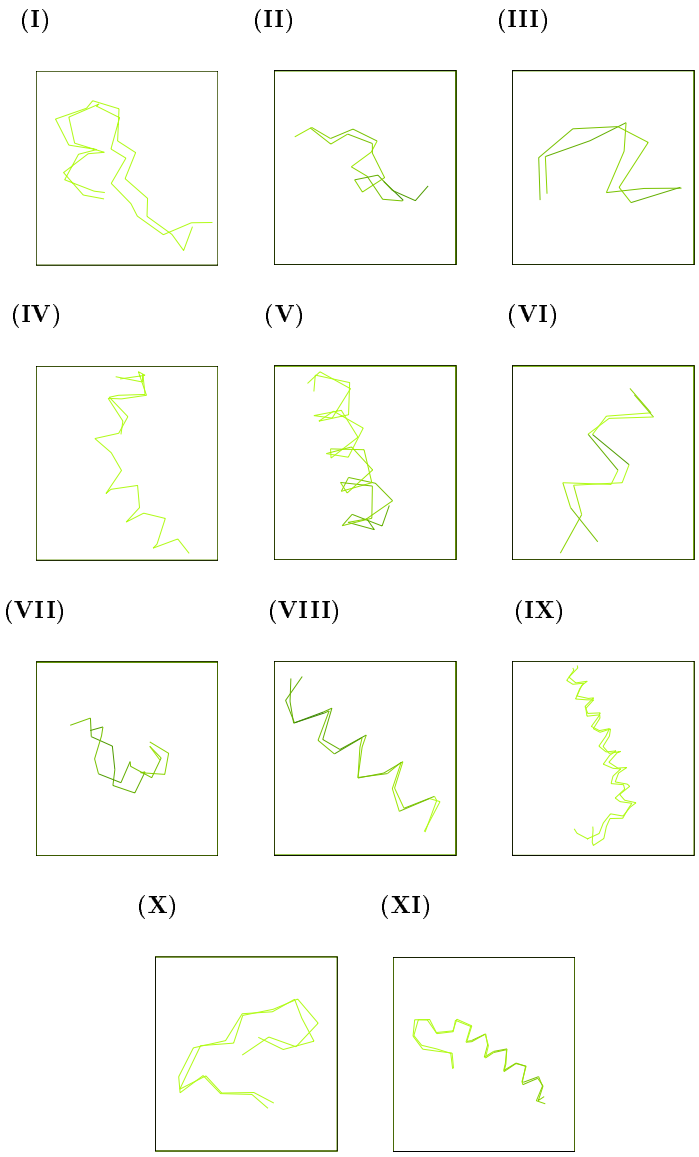


FIG. 6.11 – *Les 11 structures communes trouvées entre les cytochromes* $P450_{BM3}$ *et*$P450_{terp}$.

numéro	P450 $_{terp}$			P450$_{BM3}$			résidus	taux d'identité PBs %	taux d'identité acide aminé %	RMSD $\AA$
I	47	-	64	45	-	62	18	88.8	0.0	1.6
II	98	-	109	81	-	92	12	58.3	8.0	2.6
III	105	-	112	91	-	98	8	75.0	0.0	1.7
IV	120	-	141	106	-	127	22	95.0	18.0	0.7
V	151	-	169	139	-	157	19	89.5	0.0	1.8
VI	175	-	183	178	-	186	9	88.8	0.0	1.1
VII	169	-	180	192	-	203	12	75.0	25.0	3.6
VIII	209	-	225	201	-	217	17	94.1	6.0	0.7
IX	246	-	285	243	-	282	40	90.0	30.0	1.9
X	325	-	339	340	-	354	15	73.3	20.0	1.8
XI	369	-	396	392	-	419	28	92.8	25.0	0.7

TAB. 6.2 – *Les 11 structures locales communes avec leurs positions dans les deu protéiques et en acides aminés, le RMSD entre les deux structures, l'équivalen Haseman et collaborateurs [80]. / désigne une structure non trouvée dans les C 4,* $^{(-)}$ *le fragment* ***X*** *contient le CSB 10 et la fin du CSB 9,* $^{(a)}$ *le label n'est proximité immédiate (moins de 50 résidus),* $^{(b)}$ *il n'existe pas dans un des deux cy cytochromes et à une distance importante (plus de 50 résidus).*

hybride en utilisant la même métrique que précédemment. La figure 6.9 donne les positions des deux protéines sur la protéine hybride. Les positions sont le plus souvent continguës, effet de l'apprentissage de la protéine hybride. Un dotplot est ensuite calculé entre les deux protéines recodées par la protéine hybride. Les figures 6.10a à 6.10c montrent les dotplots obtenus pour des longueurs de fenêtres de filtrage G égal à 4, 6 et 15. Le tableau 6.2 récapitule les 11 couples de structures locales notées en chiffres romains avec leurs positions dans les deux cytochromes, leurs longueurs, leurs *RMSd*s respectifs, et, les correspondances de ces fragments par rapport aux travaux précédents de Haseman et collaborateurs [80] et de Jean et collaborateurs [94]. La figure 6.11 montre la superposition de ces 11 structures locales.

6.4.3 Analyse des résultats

Jean *et al.* avaient trouvé 15 CSBs notés de 1 à 13. Les CSB 2A-2B et 12A-12B étaient clairement liés mais n'avaient pas pu être déterminés automatiquement par leur méthode. Avec notre approche, 11 des 15 CSBs ont été retrouvés. Les CSBs 6, 11 et 13 n'ont pas été détectés du fait de leur faible taille (12, 8 et 9 résidus). Utilisant des séries de 10 blocs protéiques de longueur 5 C_α, soit 14 C_α au total, ce type de fragments est considéré comme trop petit. Le CSB 8 est plus long avec ses 17 résidus, mais n'a pas été détecté non plus. Ce dernier point est discuté plus bas. Les structures locales **I**, **IV**, **V**, **VI**,**VII** et **IX** ont la même longueur que leurs CSBs correspondants.

Quelques structures locales ont des différences notables en comparaison des précédents travaux. Ce sont les fragments **II**, **IV**, **V**, **VI** and **VII** par rapport au travail de Haseman et collaborateurs et les fragments **II**, **VII**, **X** et **XI** par rapport au travail de Jean et collaborateurs. Le taux de d'identité des blocs protéiques (cf. tableau 6.2, cinquième colonne) donne une idée de la proximité entre les structures locales. Ils dépassent globalement 73 %, sauf pour la structure locale **II** avec un taux de 58,3 %.

structure locale II: La structure locale **II** n'est pas retrouvée dans les CSBs et possède un *RMSd* important. Toutefois, ce type de structure locale est pratiquement équivalente entre les deux cytochromes, surtout dans la partie N-terminale. Un seul motif caractéristique est retrouvé dans ce fragment, une hélice 3_{10} helix, notée b, trouvée exclusivement dans le cytochrome $P450_{terp}$ [80].

structure locale IV : La structure locale **IV** est composée d'hélices α et 3_{10}, toutefois, la classification en structures secondaires note que l'hélice α C' n'existe pas dans le cytochrome $P450_{BM3}$ et les deux hélices 3_{10} *b* and *c* sont disposées à des positions distinctes. Cette différence de classification n'empêche pas un *RMSd* faible de 0,7 Å.

structure locale V : La structure locale commune **V** a un *RMSd* de 1.8 Å entre les cytochromes 450_{terp} et $P450_{BM3}$. Cette valeur plus importante est due à l'absence d'une hélice π, notée E', dans le cytochrome $P450_{BM3}$. Cette zone se retrouve incluse dans une hélice α, notée E. La différence est ponctuelle, mais fait augmenter le *RMSd*. Toutefois, il faut noter que ces structures locales possèdent le même type de repliement.

structure locale VI : Elle possède une hélice 3_{10}, notée *c*, qui se trouve incluse pour le cytochrome $P450_{BM3}$ dans le fragment **IV**.

structure locale VII : Elle correspond, comme la structure locale **VI**, au CSB 4. Toutefois sa localisation est distincte au niveau du cytochrome $P450_{BM3}$. Le *RMSd* est de 3,6 Å contre 1.1 Å précédemment. Les raisons de cette différence s'explique par le taux plus faible d'identité des blocs protéiques (75,0% contre 88,8% pour le fragment **VI**). Le CSB est composé principalement d'une hélice, notée F [80], qui présente des longueurs fort différentes selon les cytochromes, et surtout, c'est le seul CSB à avoir été sélectionné manuellement par Jean et collaborateurs.

structure locale X : La structure locale **X** inclus le CSB 10 et la fin du CSB 9.

structure locale XI : La structure locale **XI** inclus les CSBs 12A (une poche cysteique [80]) et 12B (une hélice α notée L [80]) qui avaient été décrits séparément, mais sont vraiment liés comme le montre le *RMSd* de 0,7 Å.

Ces résultats démontrent bien l'intérêt de la protéine hybride dans l'extraction de structures locales stucturellement similaires entre deux protéines. Un simple dotplot n'utilisant que les blocs protéiques ne permet pas une extraction aussi simple et rapide, du fait de la répétitivité des structures secondaires. Le bon résultat obtenu est dû au principe de l'apprentissage de la protéine hybride qui apprend des fragments protéiques et regroupent les fragments similaires au même site.

Un point à noter est que les 11 paires de structures locales n'ont jamais un taux d'identité de blocs protéiques de 100 %, mais varient entre 73,3 % et 95,0 % avec l'exception de la paire **II** avec 58.3%. Le taux d'identité de séquence en acide aminés est faible comme attendu (inférieur à 30%). Les différences majeures entre les différentes méthodes concernent les zones les plus

variables comme la structure locale **II**. Les divergences avec la classification des structures secondaires viennent du fait que cette classification place des petites structures un peu partout, sans tenir compte de la modularité des cytochromes. Par exemple, l'hélice 3_{10} *c* notée en position [173-176] pour le cytochrome $P450_{terp}$ est en position [109-114] pour le $P450_{BM3}$. De la même manière la structure locale **V** est composée d'un court feuillet β, une hélice π et d'une hélice α de 12 résidus pour le cytochrome $P450_{BM3}$. Pour le cytochrome $P450_{terp}$, le même feuillet β est présent, mais l'hélice π est alors incluse dans une hélice α de 12 résidus. Le CSB 8 n'a pas été trouvé alors qu'il est d'une taille conséquente et possède un *RMSd* correct de 1,2 Å. Cet oubli vient de la métrique du filtrage du dotplot qui est en tout ou rien. 3 sites du CSB 8 sur la protéine hybride sont différents entre le cytochrome $P450_{terp}$ et $P450_{BM3}$ et donc n'ont pas été pris en compte. Une amélioration du filtrage serait nécessaire.

6.4.4 Conclusion

Dans un premier temps, nous avons vu que la MPH permet l'obtention d'une protéine hybride qui approxime correctement un des fragments protéiques. A chaque site, un nombre important de structures similaires sont présentes. Ainsi, cette méthode permet de réduire la taille d'une base de données structurales protéiques et donc est utile dans une approche de type enfilage. Certaines améliorations sont possibles. Concernant les rares sites ayant un *RMSd* plus important que la moyenne, le rôle de la taille L des séries de blocs étant prépondérant, sa diminution pourrait améliorer la reconnaissance d'homologie structurale. Une autre optique possible est l'utilisation de fragments de tailles variables, comme l'ont fait Bystrofff et Baker[19], mais l'utilisation finale de la protéine hybride serait plus complexe. L'intérêt principal de cette méthode est le maintien de la continuité entre les fragments protéiques et permet d'observer des structures continues avec des compositions en acides aminés intéressant. L'exemple des deux cytochromes P450 montre un exemple d'application à la modélisation locale par homologie. Cette approche de la protéine hybride peut donc être fort utile dans une méthode de prédiction ou de modélisation moléculaire. De plus, la recherche de zones structuralement homologues est particulièrement simple et rapide car elle ne nécessite pas de méthode d'optimisation pour rechercher les homologies, le codage en blocs protéiques et le positionnement dans la protéine hybride suffisent.

6.5 Application à l'étude des relations séquence-structure

6.5.1 Objectif

Les protéines se replient suivant un nombre limité de conformations [70], toutefois la complexité et le nombre important de paramètres physico-chimiques, cinétiques, dynamiques et stèriques rentrant en jeu rendent la prédiction de la structure tridimensionnelle d'une protéine difficile. L'augmentation des bases de données génomiques rend la compréhension de ce problème encore plus crucial aujourd'hui.

Dans le second chapitre, nous avons utilisé une base de données structurales pour définir un alphabet structural, puis nous avons mis en relation les séquences et les blocs protéiques pour établir une stratégie de prédiction. Nous sommes donc passés de la structure à la séquence pour repasser à la prédiction soit en résumé:

$$\text{Structure} \rightarrow \text{Séquence} \rightarrow \text{Structure}$$

Avec la protéine hybride construit à partir des blocs protéiques (cf. paragraphe 6.3) nous avons caractérisé la "dépendance floue" entre la séquence et la structure locale du squelette protéique, mais d'abord en apprenant sur la structure et en déduisant ensuite les caractéristiques en termes de composition en acides amimés.

L'objectif de cette présente étude est d'analyser les dépendances entre l'information structurale et l'information séquentielle, cette dernière étant moins conservée au cours de l'évolution.

6.5.2 Principe

La protéine hybride présentée dans cette partie vise à apprendre à la fois la séquence et la structure des protéines, et donc permet d'étudier la répartition de l'information structure + séquence. Les données sur la séquence ont été recodées en fonction de critères physico-chimiques (volume, charge, hydrophobicité) et les données structurales avec l'aide des angles dièdres ϕ et ψ. L'analyse de la répartition le long de la protéine hybride en relation avec la nature du bloc structural permet d´affiner le rôle de certains acides aminés dans les structures secondaires et des régions flanquantes. L'étude aboutit ainsi à un concept de modèle flou entre la séquence et la structure, une séquence n'étant pas associée à un seul type locale de repliement, mais à plusieurs, et inversement un repliement local est catégorisé avec des séquences différentes [36].

6.5.3 données structures et données séquences

La base de données est celle préalablement utilisée, composée de 342 protéines ayant moins de 25% d'identités [84, 83]. Les protéines ont été découpées en fragments de $M = 5$ acides aminés consécutifs, ce qui donnent un base de données contenant 86980 fragments. Chaque acide aminé a été recodé selon trois variables :

* l'hydrophobicité suivant l'échelle de Kyte et Doolitle [115].
* le volume de la chaîne latèrale suivant l'échelle de Zamyatin [207].
* la charge avec trois niveaux : -0.5 attribué à K, R et H, +0.5 à D et E, et 0 aux autres acides aminés.

Les deux premières variables ont été normalisées entre -1,0 et +1,0 et sont représentées sur la figure 6.12. Les grandes catégories sont retrouvées (cf. figure 2.2). Les valeurs des trois échelles sont données dans le tableau 6.3

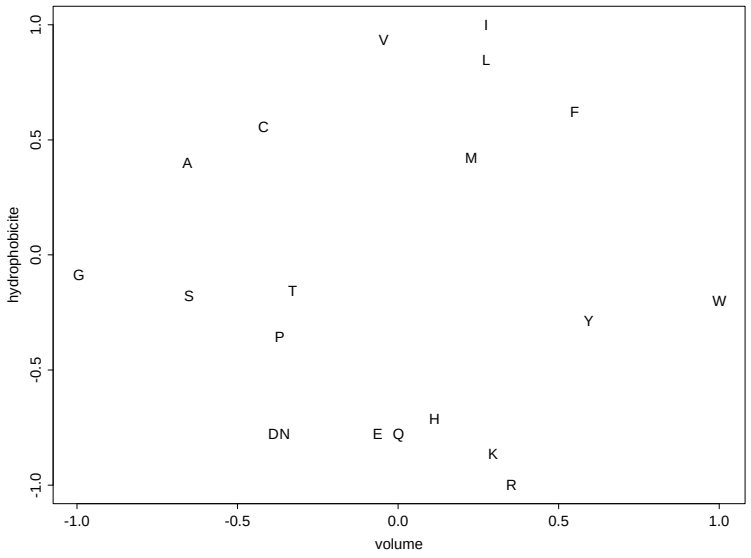


FIG. 6.12 – *Représentation de l'échelle des volumes des acides aminé de Zamyatin [207] en fonction de celle de Kyte et Doolittle [115]. Les deux échelles sont normalisées entre -1,00 et +1,00.*

La structure trimensionnelle de la chaîne carbonée associée à 5 résidus consécutifs (le résidu central étant en position s dans la séquence protéique) est caractérisée par 8 angles dièdres ψ_{s-2}, ϕ_{s-1}, ψ_{s-1}, ϕ_s, ψ_s, ϕ_{s+1}, ψ_{s+1}, ϕ_{s+2}) qui ont été normalisés entre -1,0 et +1,0. Ils ont été préalablement décalés pour les angles ϕ supérieurs à 120° de -360° et pour les angles ϕ inférieurs

acide aminé	hydrophobicité	volume	polarité
A	-0,66	+0,40	0,00
R	+0,36	-1,00	-0,50
D	-0,39	-0,78	+0,50
N	-0,36	-0,78	+0,50
C	-0,42	+0,56	0,00
E	-0,06	-0,78	0,00
Q	0,00	-0,78	0,00
G	-1,00	-0,09	0,00
H	+0,11	-0,71	-0,50
I	+0,27	+1,00	0,00
L	+0,27	+0,84	0,00
K	+0,30	-0,87	-0,50
M	+0,23	+0,42	0,00
F	+0,55	+0,62	0,00
P	-0,37	-0,36	0,00
S	-0,65	-0,18	0,00
T	-0,33	-0,16	0,00
W	+1,00	-0,20	0,00
Y	+0,59	-0,29	0,00
V	-0,04	+0,93	0,00

TAB. 6.3 – *Ensemble des variables avec les échelles normalisées de l'hydrophobicité [115], du volume [207] et de la polarité.*

à -120^o de +360^o. Chaque fragment de 5 résidus est donc défini par un vecteur $\mathbf{V}$ ayant $m = 23$ composantes (15 pour la séquence et 8 pour la structure), toutes comprises dans l'intervalle [-1,0:+1,0].

6.5.4 Apprentissage de la protéine hybride

Dans notre étude, la protéine hybride correspond à une succession de L fragments d'une longueur $M = 5$ résidus, chacun caractérisé en termes séquence-structure par un vecteur de m composantes (ici, $m = 23$). Elle est donc symbolisée par une matrice de dimension L x m. Le principe de base de MPH est d'apprendre "au mieux" l'ensemble de la base de vecteurs (au nombre de 86 980) par cet hybride de L vecteurs. L'apprentissage est similaire à celui d'une carte auto-organisée de Kohonen ("Self-Organizing Map" ou SOM [108, 109]). Cependant, dans notre cas, l'apprentissage est monodimensionnel et la diffusion de l'information le long de l'hybride n'est pas réalisée artificiellement. Elle est implicite car plusieurs vecteurs successifs sont utilisés à la même étape de l'apprentissage. Ce point est distinct de celui de la précédente protéine

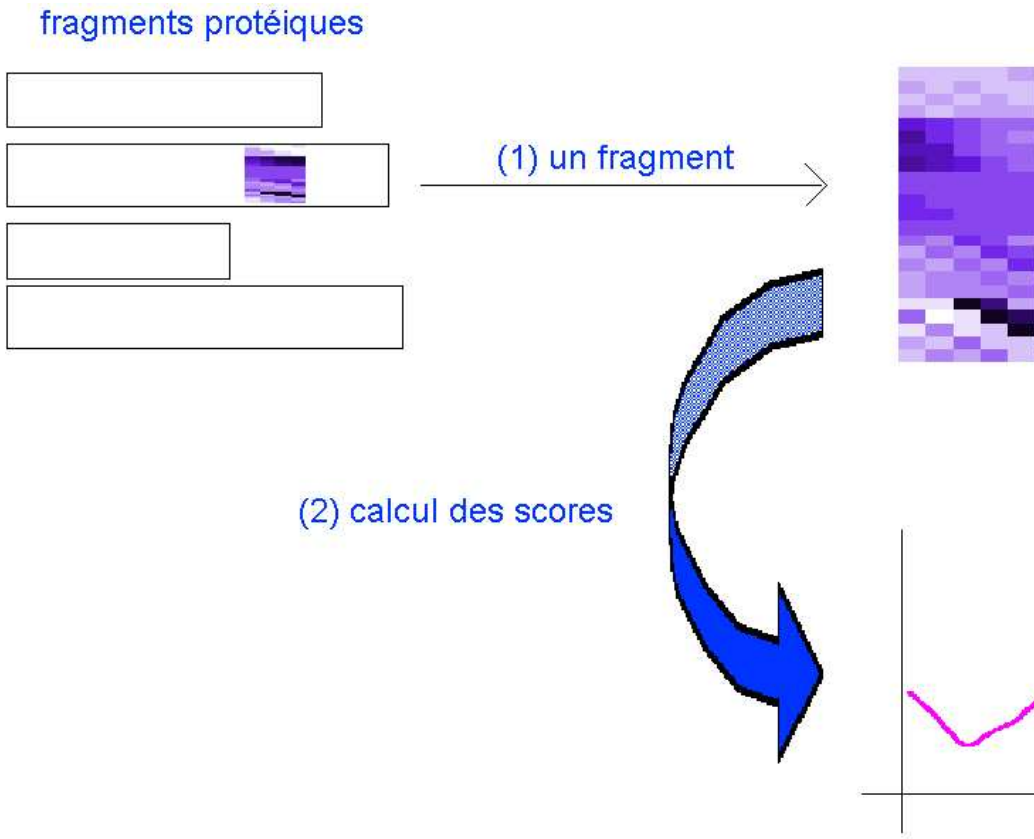


FIG. 6.13 – *Schéma représentant le principe d'apprentissage des informations séq*
fragment défini par une sous-matrice d'observations. (2) Calcul d'un score local e
(3) Détermination de la position optimale dans l'hybride en recherchant le sco
l'hybride.

hybride sur les séries de blocs protéiques où la continuité était comprise directement dans la série de blocs protéiques. Un vecteur d'observation était présenté avec un seul type de blocs pour chaque position alors qu'ici une sous-matrice de f vecteurs est utilisée ($f = 5$ dans cette étude). En effet, nous présentons f vecteurs consécutifs à l'hybride pour effectuer un apprentissage en continu à la fois de la séquence et de la structure. La méthode schématisée dans la figure 6.13 se décompose en trois étapes :

(i) Initialisation de la protéine hybride : on effectue un tirage au hasard de L vecteurs dans les protéines codées.

(ii) Apprentissage séquentiel des matrices d'observations :

(1) on tire un fragment avec son environnement de taille f de la base de données. Il est défini par une sous-matrice $\boldsymbol{V}$ de f vecteurs de taille m.

(2) Pour chacune position p de l'hybride un score de dissemblance $\boldsymbol{S}(p)$ (une distance euclidienne) est calculé entre la sous-matrice $\boldsymbol{V}$ et celle $\boldsymbol{W}(p)$ de même taille prise dans l'hybride. Ainsi un profil de scores est établi le long de l'hybride. (3) On recherche alors le score minimum S_{min} et la position $p^* = argmin\{\boldsymbol{S}(p)\}$ associé à la plus forte ressemblance entre le fragment observé dans une protéine et celui dans la protéine hybride. (4) Ayant localisé le fragment, on va donc modifier légèrement le contenu de la sous-matrice $\boldsymbol{W}(p)$ pour qu'elle ressemble davantage à celle présentée, soit $\mathbf{V}$. La transformation est définie par l'équation :

$$\boldsymbol{W}(p) \rightarrow \boldsymbol{W}(p) + (\boldsymbol{V} - \boldsymbol{W}(p)).\alpha(n)$$

et

$$\alpha(n) = \alpha_0/(1 + n/N)$$

n désignant le nombre de sous-matrices présentées à l'hybride, N le nombre total de sous-matrices de la base de données et α_0 le coefficient d'apprentissage initial. Le coefficient d'apprentissage α(n) décroit au cours de l'apprentissage. Ayant modifié l'hybride, on passe au fragment suivant jusqu'à traiter complétement la base.

(iii) Renforçage de l'apprentissage : on effectue un certain nombre C de cycles d'apprentissage de la base en recommençant l'étape (ii). Cette relecture des informations permet de renforcer l'apprentissage en regroupant progressivement les blocs semblables.

6.5.5 Résultats

La figure 6.14 donne le résultat de l'apprentissage effectué avec un coefficient d'apprentissage $\alpha_0 = 0{,}03$ pour $C = 20$ cycles pour une protéine hybride de longueur $L = 25$. La séquentialité des fragments est visible, toutefois le vecteur caractéristique du fragment en position p ne présente pas exactement le même vecteur décalé du fragment en position *(p - 1)*. D'après les variations en niveau de gris, les variables hydrophobicité et angle dièdre ϕ semblent jouer un rôle important, et à un moindre niveau, le volume et l'angle dièdre ϕ. La charge joue un rôle mineur, cela peut s'expliquer par le nombre faible de chargés par rapport au nombre total de résidus ou par un problème d'étalonnage des variables. Les 25 positions présentent globalement des caractéristiques différentes. Cependant une classification des 25 positions par une classification hiérarchique (paquage *hclust* du logiciel *S-Plus*) montre la constitution de deux groupes homogènes distincts ayant comme frontières les deuxième et treizième positions. Après apprentissage, le nombre de fois où une position de l'hybride est sélectionnée a été calculée. La répartition des observations est relativement homogéne le long de l'hybride, les nombres variant entre 2950 et 4500 observations.

6.5.6 Correspondance entre la protéine hybride et les blocs structuraux

La figure 6.15 donne la composition en acides aminés du résidu central des 25 fragments (ce n'est qu'une information partielle) sur sa partie gauche, et les fréquences relatives des fragments pour chaque bloc structural. Afin de simplifier cette figure, seuls les groupes dont le nombre de fragments attribués était supérieur à 100 et dont la fréquence dans le bloc protéique était supérieure à 4 % (i.e. 1/L) ont été mis. Il s'avère que seulement 15,6 % des fragments de la base n'ont pu être attribués. Les constations déduites de l'étude des figures sont les suivantes :

(i) une dépendance forte entre les fragments de l'hybride et les blocs structuraux; des positions 2 à 13, les blocs qui concernent les hélices α et leurs régions flanquantes sont

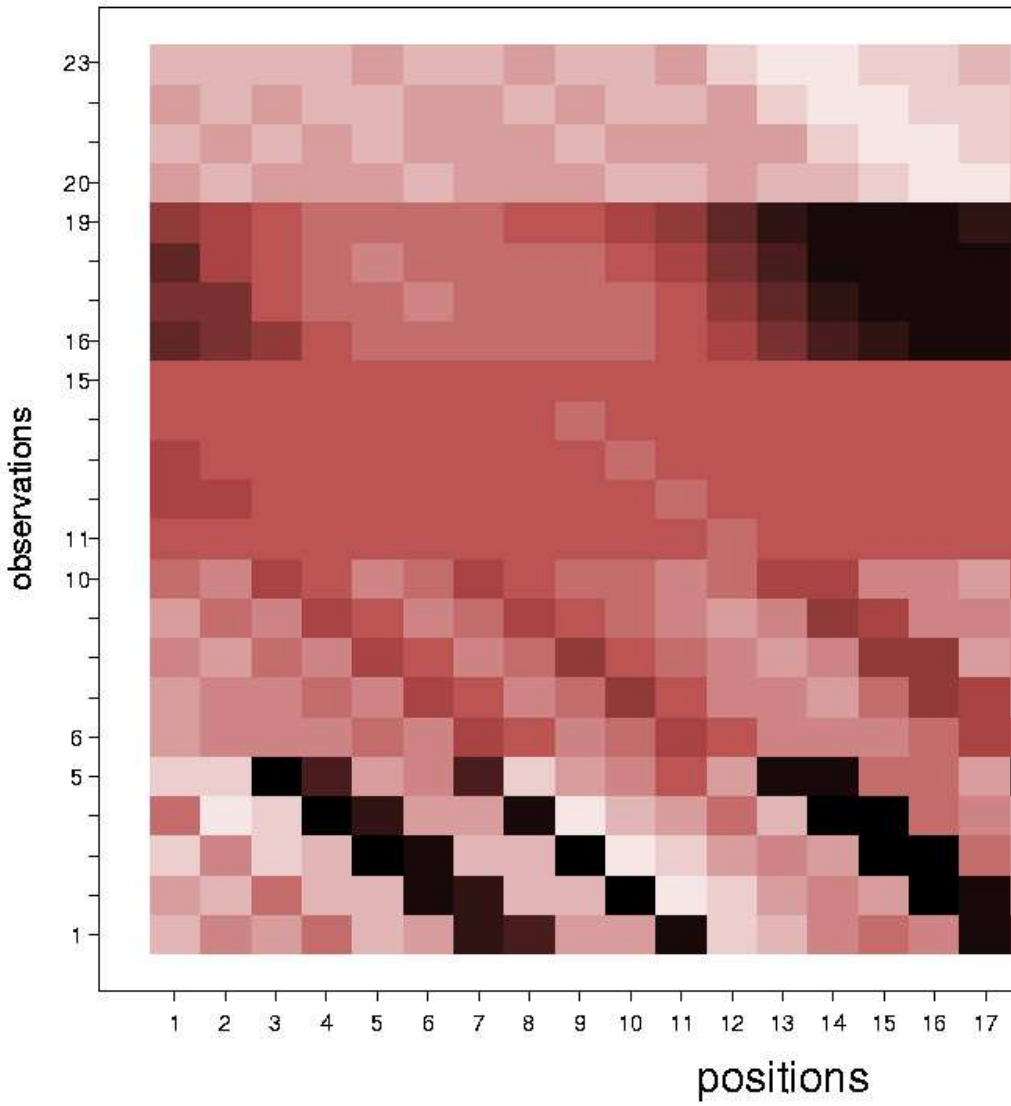


FIG. 6.14 – *La protéine hybride après apprentissage. La protéine hybride e*
vecteurs de 23 observations ont été déterminés par l'apprentissage. En ordonné
[1:5] l'hydrophobicité, [6:10] le volume, [11:15] la charge, les angles dièdres [16:1

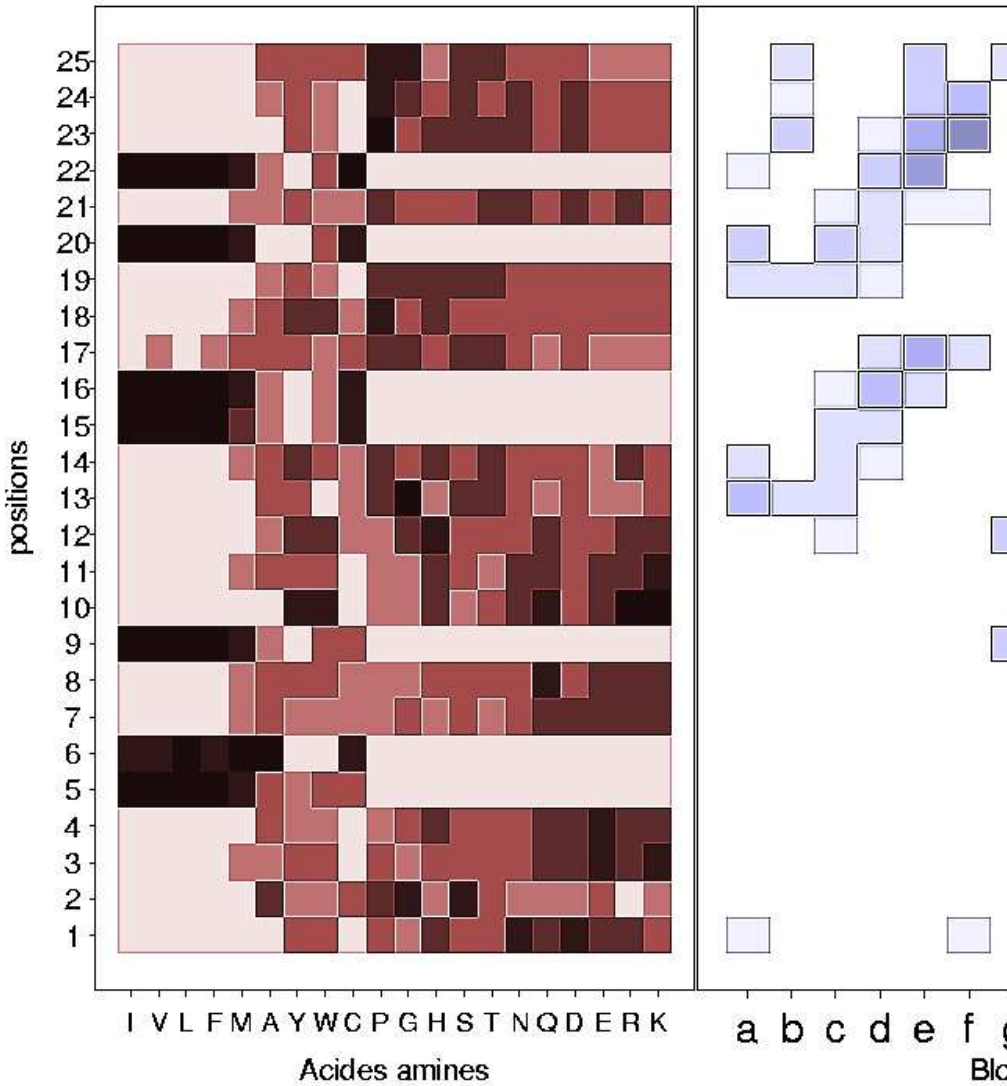


FIG. 6.15 – *Correspondances entre la protéine hybride et les blocs structuraux long de la protéine hybride uniquement pour l'élément central des fragments. (b) structuraux.*

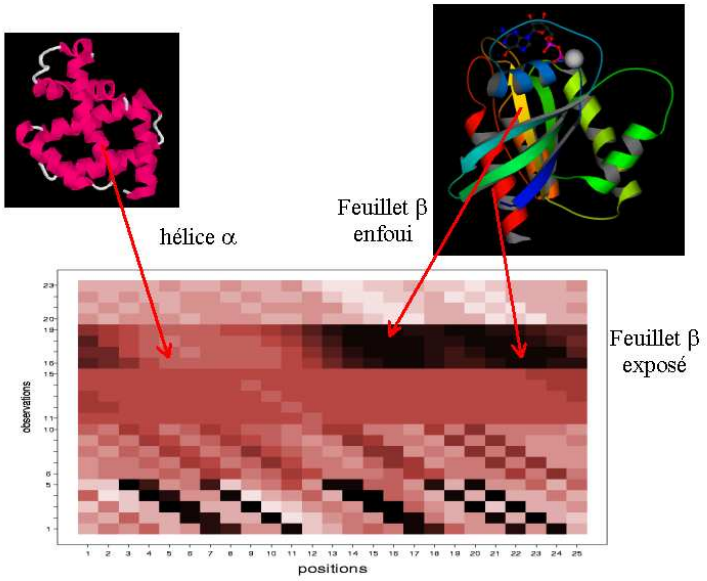


FIG. 6.16 – *Positionement des hélices et des feuillets dans la protéines hybrides.*

retrouvées, et de 13 à 25 (cf. figure 6.16), les feuillets β, leurs régions flanquantes et les boucles. Ces dernières sont localisées dans les positions 1, 12-13, 17-19 et principalement 22-25. On note des sur-représentations des Glycines et des Prolines dans ces zones en association avec les blocs protéiques h, i et j.

(ii) la séquentialité des blocs structuraux et des fragments dans l'hybride se retrouvent dans la figure de droite. On observe deux tendances dans les feuillets β, l'une aux positions de 13 à 19, et l'autre de 20 à 25. Le BPd contient deux hydrophobes consécutifs dans le premier type qui peut correspondre à une hélice enfouie, et deux hydrophobes séparés par un hydrophile dans le second type qui correspond à un feuillet dont l'un des brins est enfoui et l'autre exposé (cf. figure 6.16).

(iii) certains blocs protéiques se retrouvent majoritairement en certaines positions de l'hybride. Par exemple, BP l (extrémité N-terminale d'une hélice α) est localisé en positions [1:4] où apparaissent des résidus chargés, BPm (région centrale d'une hélice α) en positions [4:11] dans lesquelles des résidus hydrophobes (Isoleucine, Valine, Leucine, Phénilalanine, Méthionine et partiellement Alanine) et où un motif hydrophobe est retrouvé (i, i+1, i+4), une partie de l'hélice est enfouie.

(iv) les positions centrales (5, 6, 15, 16, 20 et 22) présentent un caractère hydrophobe très marqué. La Cystéine occupe aussi les mêmes positions, cependant à un degré moindre

dans d'autres positions,

(v) le rôle des résidus chargés est moins précis alors qu'ils devraient intervenir dans les régions flanquantes des hélices et des feuillets.

6.5.7 Conclusion

La répartition des Blocs Protéiques montre ainsi une forme de regroupement que l'on pourrait simplement comparé à l'alphabet structural classique à 3 états (hélices, feuillets et boucles). Un simple comptage des 3 types de structures donne une information évidemment corrélée à celle obtenue dans l'hybride. Toutefois, cette information ne tient aucun compte de l'aspect séquentiel. La correspondance avec l'alphabet structural à 16 états permet au niveau de la structure d'avoir une information bien plus précise. On peut ainsi voir la spécificité de l'entrée dans l'hélice α (autour de la position 3 principalement) sur le plan de la composition en acides aminés, de même en sortie d'hélice (positions 8-11). L'hybride permet de tenir compte de l'hétérogénéité de longueur inhérente aux structures répétitives. De plus l'utilisation des structures secondaires se heurte à l'établissement toujours arbitraire de règle d'attribution à l'une des 3 classes [31, 33]. Avec les feuillets β, l'exemple est encore plus frappant, l'apprentissage ayant permis d'obtenir deux types distincts de feuillets (positions 14-17 et positions 19-23), de longueurs et surtout de compositions fort distinctes.

La protéine hybride permet, en apprenant structure et séquence de concert, d'avoir une compression de ces données en un nombre fini d'états, où les deux types d'information sont combinés. L'utilisation des 16 BPs, outil intéressant pour analyser les structures, permet de distinguer de facon fine les divers états de la structure. En conclusion, la correspondance entre la succession des fragments de 5 résidus dans la protéine hybride et les différents types de Blocs Protéiques d'autre part a permis de mettre en relief le concept de relation séquence-structure "floue". C'est à dire qu'une chaîne d'acides aminés est associée à une distribution de patterns structuraux (i.e. les BPs) et inversement. Cela implique que sur le plan d'une prédiction de la séquence vers la structure, certains blocs protéiques doivent être considérés comme équivalents.

Un tel concept devrait être pris en compte dans la construction de modèles structuraux protéiques que ce soit par une stratégie d'enfilage (threading) ou une modèlisation *ab initio*. Cette méthode peut aussi de servir de validation lors de l'élaboration de modèles structuraux

complets, à la manière de ProCheck avec la carte de Ramachandran. La figure 6.15 de gauche pourrait aussi servir à l'analyse des relations entre séquence et structure à l'intérieur de familles de protéines. Toutefois, des améliorations restent à effectuer, comme par exemple, changer d'échelles pour décrire les acides aminés.

6.6 Comparaison entre les cartes de Kohonen et MPH

Juste avant de conclure ce dernier chapitre, je désire récapituler les différences et les ressemblances méthodologiques existantes dans la méthode que nous avons mis au point et les cartes de Kohonen.

Les cartes de Kohonen sont des outils puissants d'analyse (cf. Annexe 1), elles sont donc le plus souvent bidimensionnelles, MPH est lui unidimensionel. L'apprentissage s'effectue dans les deux cas par un processus itératif, avec un coefficient d'apprentissage faible qui permet une modification locale limitée décroissant avec le nombre d'observation présenté.

La plus grande distinction entre les deux méthodes concerne la diffusion de l'information. Dans une carte de Kohonen, une diffusion contrôlée sur les neurones voisins existe. Ainsi, quand un neurone U_{i-x} est modifié de α %, les neurones contigus U_{i-x} et U_{i+x} seront modifiés de α/τ % ($\tau > 1$). Cette diffusion de l'information permet de rassembler dans une zone des neurones proches. Pour MPH, aucune diffusion n'est effectivement effectuée, elle est remplacée par un processus d'empilage. Les neurones U_{i-x} et U_{i+x} vont apprendre une partie de l'information, un décalage s'effectue naturellement, les observations prises avec leur environnement et les neurones contigus au vainqueur vont apprendre cet environnement. Ainsi, une continuité s'intaure entre les neurones.

Les neurones obtenus ne sont donc pas indépendant comme dans les SOMs mais "séquentiels", un neurone U_i est conditonné par ses neurones voisins U_{i-x} et U_{i+x}.

6.7 Conclusion

La méthode de la protéine hybride (MPH) est un modèle flou de compaction de la structure locale des protéines. La protéine hybide lors de la première étude a permis l'apprentissage d'une base de données structurales recodées en blocs protéiques. Elle est donc composée d'une série

de lois de probabilité d'observation des blocs. L'apprentisage est un processus qui consiste pour chaque série de blocs à trouver un site qui lui est très proche et à le modifier faiblement. Ce processus nous a permis de cosntruire une protéine hybride qui possède en chaque position une série de blocs bien déterminée, et, ces séries quoique parfois différentes possèdent entre elles une déviation quadratique moyenne trés acceptable de 3,14 Å pour des longueurs de 10 C_α. En plus de l'analyse des structures et des acides aminés qui leurs sont associés, l'utilisation de la protéine hybride dans la recherche de zone d'homologie structurale entre deux cytochromes P450 montre une efficacité évidente. L'utilisation conjointe de cette information structurale avec une recherche sur la séquence dans une méthode de modélisation par homologie devrait-être envisagée. De même, la compaction de la base de données en une centaine de prototypes peut permettre une diminution du nombre de candidats à tester dans une approche de type enfilage. Enfin, MPH a montré ses capacités dans l'établissement des liens entre la séquence et la structure. Cette dernière approche peut servir à de nombreuses méthodes comme validation de la compatibilité entre un modèle construit et la séquence, ou encore dans l'analyse d'incompatibilité de repliements dans une méthode de repliements *ab initio*.

Chapitre 7
Conclusion et perspectives

a. Conclusion générale Lors de cette thèse, j'ai mis au point une nouvelle méthodologie pour concevoir des petits prototypes protéiques, les **blocs protéiques** (BPs). Après avoir conçu plusieurs séries ayant des nombre de BPs variables, nous avons conservé un **alphabet structural** composé de 16 BPs. Il permet à la fois une approximation correcte de la structure tridimensionnelle des protéines et une prédiction acceptable. La qualité structurale des blocs a été vérifiée, ils sont à la fois bien distincts les uns des autres et permettent une assignation non ambiguë. Les BPs sont, d'un point de vue structural, hautement spécifiques.

Le choix de 16 BPs est adéquat pour analyser la structure tridimensionnelle des protéines. Deux BPs sont spécifiques des parties régulières des protéines, les autres blocs permettent ainsi de bien analyser les extrémités N- et C-terminales de ces structures. Quatre PBs sont même totalement localisés dans les parties "boucles". La correspondance avec les structures secondaires classiques sont bien retrouvées, mais les PBs permettent une analyse beaucoup plus poussée de l'ensemble des structures protéiques.

La **prédiction de la structure 3D** en termes de blocs protéiques à partir de la séquence **par une approche bayésienne** atteint un taux de 30,0 % avec une fenêtre initiale de 5 résidus. L'agrandissement de cette fenêtre à 15 résidus permet un gain de 4,4 %. La mise au point d'une méthode de séparation des séquences (**1 repliement local $\rightarrow$ n séquences**) en classes, dites **familles séquentielles**, pour les blocs les plus fréquents permet un taux final de prédiction de 40,7 %.

La spécificité séquentielle importante des blocs a montré, en plus, de ce taux correct de prédiction, que la majorité des blocs réels se trouvait parmi les plus probables. Ce fait a permis, en plus de la méthode bayésienne de prédiction, de mettre en œuvre deux stratégies basées sur

cette information (**1 séquence → n repliements locaux**): (i) une **stratégie globale** qui permet de déterminer le nombre de blocs protéiques à conserver en chaque site pour atteindre un taux de prédiction donné, (ii) une **stratégie locale** qui permet de délimiter un certain nombre de zones ayant un taux de prédiction donné en prenant un nombre fixe de blocs protéiques.

La recherche des motifs de 5 BPs les plus fréquents dans la base de données a permis de construire un **réseau** qui les connectent. Plus de 83 % des sites protéiques sont contenus dans ce graphe et l'analyse de leur répartion en acides aminés montrent que ce réseau permet de voir des différences significatives selon la position du BP dans le réseau. La vérification de la stabilité structurale du réseau a donné des résultats particulièrement bons avec des *RMSd* faibles (2 Å) pour des fragments ayant des longueurs allant de 8 à 11 C_α.

Allant plus loin dans l'utilisation de l'alphabet structural, nous avons élaboré la **méthode de la protéine hybride** (MPH). Elle permet la réduction d'une base de données structurales protéiques à une centaine de fragments prototypes d'une longueur de 10 PBs chacun. La grande majorité de ces fragments est bien approximée. Leur répartition en acides aminés montre, en outre, une spécificité importante qui est, de plus, liée dans les 3 cas sur 4 des cas à une information structurale bien déterminée. MPH appliquée aux blocs protéiques permet, de plus, une **recherche d'homologie structurale** rapide et performante.

Enfin, la méthode MPH appliquée à des informations séquences et structures recodées en paramètres physico-chimiques et angulaires a montré son intérêt dans l'analyse des relations entre les deux types d'information. Les blocs protéiques montrent ici aussi leur pertinence dans le domaine de l'analyse en permettant une description bien plus complexe des repliements qu'une simple description en structures secondaires.

b. Perspectives La figure 7.1 montre la logique des recherches effectuées et futures, en traits bleu foncé les travaux déjà réalisés, et, en bleu clair différentes voies possibles pour l'établissement d'un modéle protéique. Différentes évolutions sont possibles:

- Un premier point concerne l'**homologie structurale**, quand les structures tridimensionnelles sont connues. Déjà, la méthode MPH a démontré ses possibilités. Il conviendrait, en plus, de l'information structurale, de prendre en compte la séquence qui, quoique moins conservée dans l'évolution, posséde une information d'intérêt.

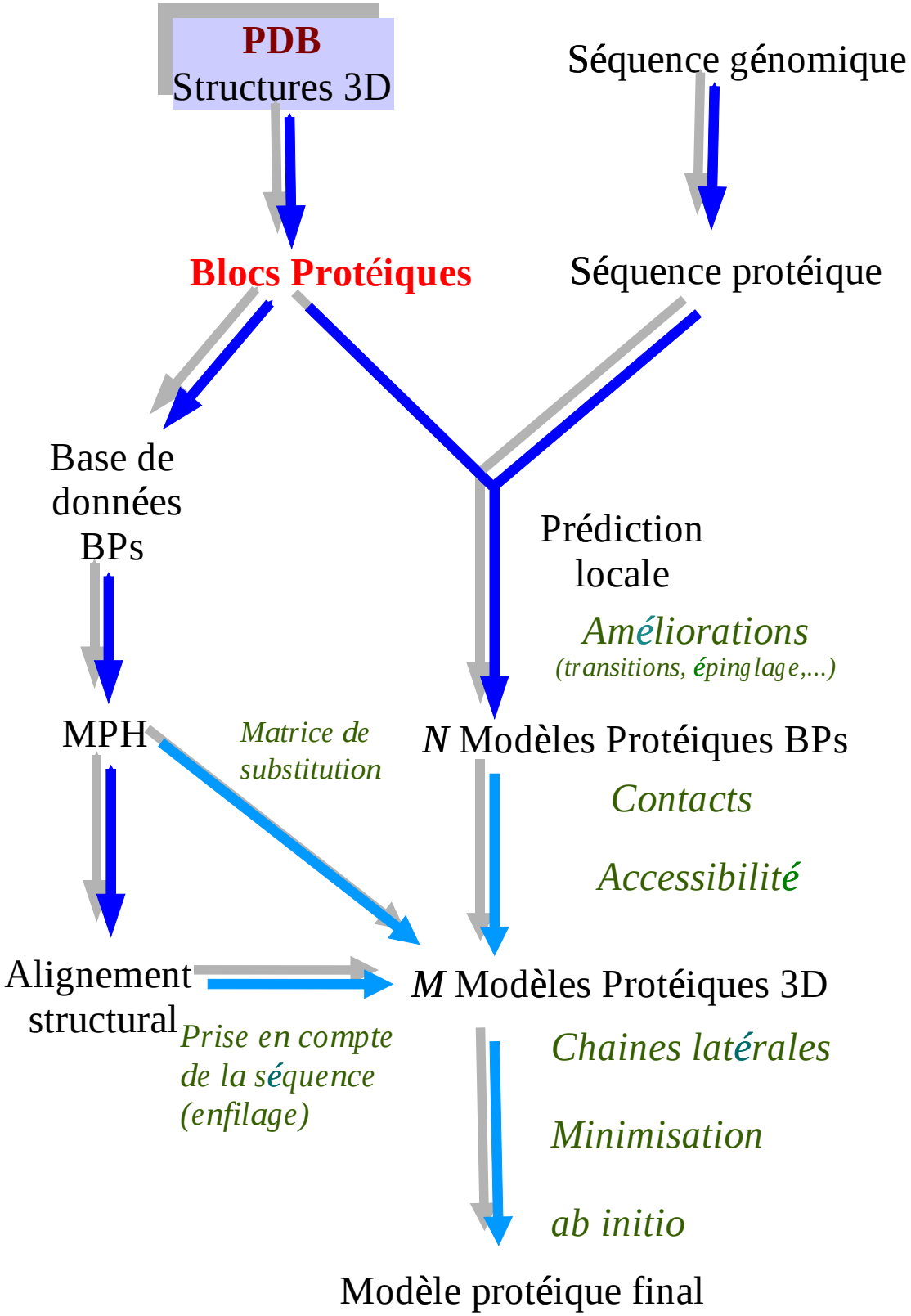


FIG. 7.1 – *Schéma récapitulatif des avancées dûes aux BPs et leur intégration dans un processus de modélisation moléculaire classique.*

- Dans la même optique, l'utilisation des blocs protéiques pourrait servir, dans une étape de prétraitement, dans une technique d'**enfilage**. En effet, observer les compatibilités entre les résultats de la prédiction locale et des structures connues issues d'une base non redondante permet de diminuer le nombre de repliements à tester.

- Un travail intéressant, touchant toujours à l'homologie structurale, est la **classfication protéique**. Ayant décrit un alphabet structural, la protéine hybride a bien montré la catégorisation de fragments protéiques en un nombre limité de classes. L'utilisation de ces observations devraient être pertinente pour reclasser les structures tridimensionnelles. Une analyse sur les familles obtenues serait bien entendu utile à la compréhension du repliement protéique.

- Un autre point concerne la **génomique structurale**. Partant d'une séquence codante, nous désirons obtenir un modèle tridimensionnel final le plus proche de la réalité. Pour cela, nous avons développé un alphabet structural et une méthode de **prédiction locale**. Cette dernière peut être améliorée par des approches classiques qui prennent en compte les alignements de séquences ou les transistions existants entre les BPs. En effet, que ce soit par le réseau sur les mots ou l'approche de la protéine hybride avec les blocs protéiques, on a observé qu'une série de BPs est suivie par un nombre limité d'autres séries bien déterminées. Actuellement, une méthode de prédiction basée sur ce principe est en fin de réalisation et montre un intérêt certain.

L'ensemble de ces résultats permet la génération de **différents modèles 3D** qu'il faut travailler. L'usage de la librairie de recodage d'angles est ainsi un premier pas, les angles dièdres étant nettement mieux approximés. De nombreux autres informations devraient être introduits pour aboutir à ce modèle, comme l'accessibilité des résidus, les contacts les plus probables entre résidus, et, le positionnement des chaînes latérales. L'utilisation comme contrainte dans un processus *ab initio* d'une manière similaire aux travaux du groupe de Baker est aussi envisageable [178]. En effet, les contraintes à longues distances ne sont pas prises en compte actuellement.

L'ensemble des travaux effectués donnent des résultats concluants. A partir d'une séquence, il est possible de délimiter des zones hautement probables et ces zones sont décrits par un alphabet structural qui permet une approximation correcte des repliements. Toutefois, des améliorations restent à faire avant d'aboutir à un modèle final.

Annexe 1 : les cartes de Kohonen

réseaux neuronaux artificiels Ce sont des modèles auto-organisés et connexionnistes. Un réseau d'auto-organisation est un réseau d'éléments de traitements simultanément actifs (nœuds et connexions). Les modèles connexionnistes utilisent une information numérique et sont des systèmes dynamiques qui effectuent des calculs analogues à ceux d'un neurone.

Un modèle connexionniste est caractérisé par trois constituants de base: un réseau, une règle d'activation et une règle d'apprentissage.

Le *réseau* est composé des nœuds (*unités*) connectés par des liens orientés (*connexions*). La règle d'activation d'un modèle connexionniste est une procédure locale que chaque nœud suit en mettant à jour son niveau d'activation en fonction du contexte d'activation des nœuds voisins. Deux aspects sont à voir à ce niveau, tout d'abord le parallélisme massif de l'activation qui implique une diffusion de l'activité et le caractère local de l'information traitée par chaque nœud.

La *règle d'apprentissage* est la propriété du réseau à changer son comportement d'après les résultats de ses activations passées. Localement, le poids de la connexion de chaque nœud est réévaluée en fonction de sa valeur actuelle et des niveaux d'activations des nœuds qu'elles connectent.

Les réseaux linéaires forment une classe particulière de modèles neuronaux (cf fig. 7.2).

On doit distinguer l'apprentissage supervisé où un règle delta qui minimise l'erreur quadratique est appliquée, le couple entrée-sortie est présenté, la réponse est donc pondérée par rapport aux résultats et l'apprentissage non supervisé où le réseau est simplement exposé aux différents exemples sans aucun type de correction.

Explication biologique du modèle de Kohonen Le modèle de Kohonen repose sur l'observation neurophysiologique que les détecteurs de caractéristiques de différentes aires du cortex sensoriel loin d'être indépendants sont en fait regroupées en carte dont la topologie est corres-

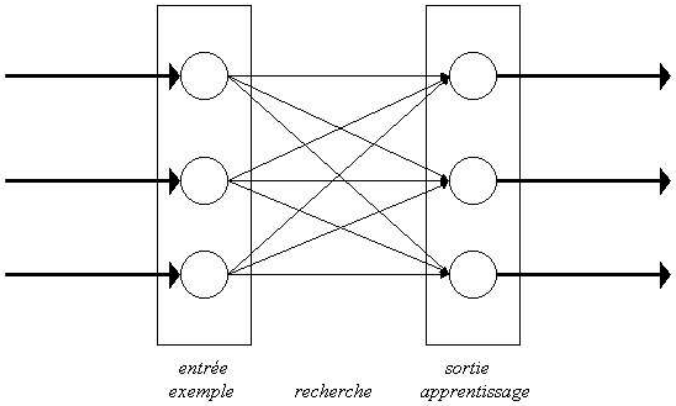


FIG. 7.2 – *Réseau à deux couches et à circulation de l'information dirigée vers l'avant.*

pondance avec la topologie spatiale. Les activités de des détecteurs de caractéristiques sont corrélées suivant la relation de voisinage pour cette topologie corticale. Il se rapproche d'un algorithme de d'apprentissage concurrentiel. Il n'est pas supervisé [108].

Principe d'apprentissage et d'assignation Une couche de neurones d'entrée définit l'espace des entrées possibles. une couche de neurones de sorties sera mise en correspondance avec l'espace des prototypes. Les poids synaptiques reliant les neurones de la couche d'entrée à un neurone de la couche de sortie définissent les coordonnées dans l'espace des entrées du prototype représenté par le neurone de sortie.

L'apprentissage tient compte de la structure topologique de la grille des entrées. L'accroissement de l'algorithme de d'apprentissage concurrentiel est appliqué au plus proche prototype du motif d'entrée mais aussi aux voisins de ce motif.

Dans un premier temps, une phase de *structuration topologique* voit la grille de la couche des sorties se positionne dans l'espace des entrées. Ses points s'ordonnent les uns par rapport aux autres.

Ensuite, une *phase de convergence* où la grille se déforme lentement en conservant sa structuration pour converger vers un échantillonnage régulier de la distribution de probabilités de l'espace des entrées.

Après apprentissage, un motif d'entrée sera représenté par le neurone dont il s'approche le plus. Par conséquent, une fois l'apprentissage achevé, les valeurs des connexions définissent un pavage de l'espace des entrées qui doit échantillonner au mieux la distribution de proba-

bilité des motifs d'entrée. La principale caractéristique est que la carte topologique obtenue n'a aucun rapport avec les dimensions de l'espace des entrées. Elle est stable, robuste et d'une représentation simple.

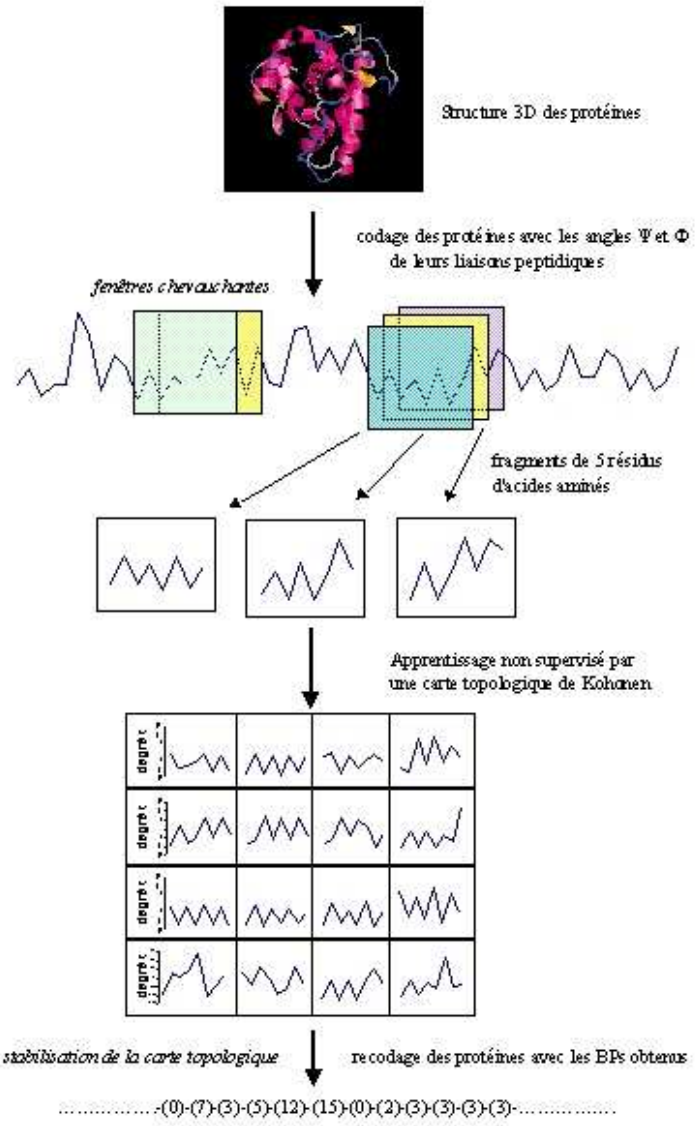


FIG. 7.3 – *Schéma récapitulatif du travail effectué par Schuchhardt et collaborateurs, avec la traduction de la structure tridimensionnelle des protéines en termes d'angles ϕ et ψ, puis leur utilisation dans une carte de Kohonen pour obtenir des blocs structuraux.*

L'intérêt des cartes topologiques est la diffusion latérale qui permet de diminuer les risques de minimum local comme dans une méthode de nuées dynamiques ou de k-means. Une carte topologique est représentée par un réseau de N neurones. Il est initialisé en tirant au hasard des fragments de la base de données pour donner des valeurs initiales aux neurones. Pour expliciter ceci, l'article de Schuchhardt et collabarateurs servira d'exemple (pour la problématique), la

figure 7.3 récapitule les étapes décrites au paragraphe 2.3.2.3.

La figure 7.4 explicite les différentes étapes pour un exemple de 16 blocs :

i)- Tirage au sort d'un fragment (vecteur d'angles dièdres)

ii)- Comparaison de ce fragment avec tous les neurones. Pour connaître le neurone le plus proche du fragment présenté, nous avons cherché le minimum de différence entre le fragment et les neurones de la carte :

$$RMSda(w^{next}, v) = \textbf{minimum}$$

avec w^{next} le vecteur de poids le plus proche de v le fragment présenté

$$RMSda(s,t) = \sqrt{\frac{1}{16}\sum_{1}^{8}(\phi_s^i - \phi_t^i)^2 + (\phi_s^{i+1} - \phi_t^{i+1})^2}$$

avec s et t représentent deux motifs structuraux distincts. Les différences angulaires se font sur 180^o au maximum.

iii)- Le neurone le plus proche (valeur de rmsda la plus faible) est modifié très légèrement pour ressembler au fragment qui lui est présenté. Les poids (vecteur de 8 coordonnées pour chaque neurone) sont pondérés à chaque présentation de fragments :

$$w^k(t+1) = w^k(t) + [v - w^k(t+1)]\nu e^{-\frac{1}{2\rho^2}(r^k - r^{next})}$$

avec ,

$$\nu = \frac{\nu_0}{1 + \frac{t}{\theta}}$$

$$\rho = \frac{\rho_0}{1 + \frac{t}{\theta}}$$

Avec θ le nombre total de motifs à passer, t le nombre de motifs déjà passés, ν le coefficient d'apprentissage pris classiquement entre 0,01 et 0,02 ρ est un coefficient qui définit la distance de propagation (cf. iii bis) autour du neurone le plus proche wnext et r l'amplitude des modifications autorisées.

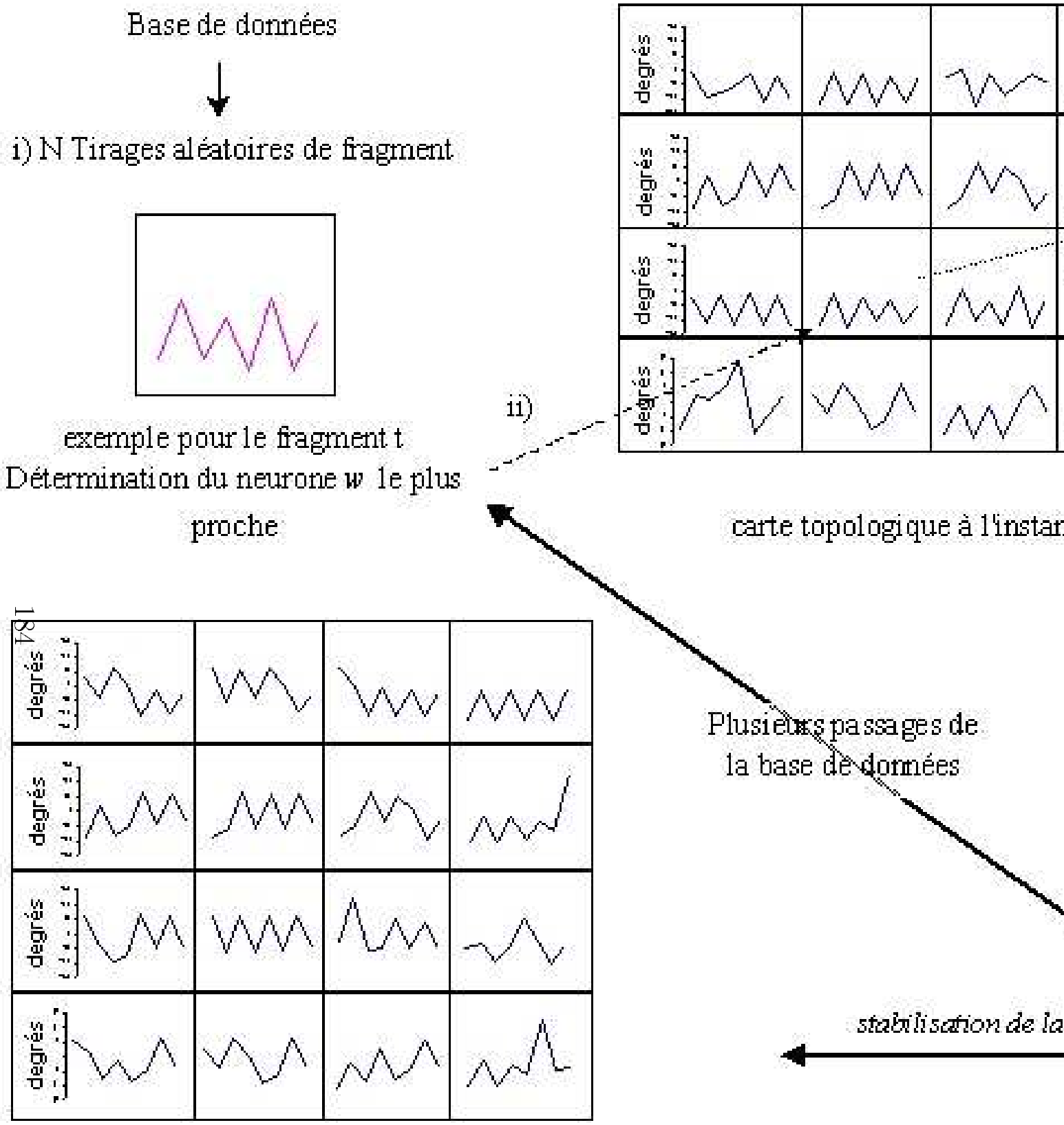


FIG. 7.4 – *Principe de l'apprentissage d'une SOM. (i) Un fragment est tiré al chaque neurone de la carte topologique, (iii) le neurone vainqueur est sélection sont modifiés plus faiblement, le processus est répété jusqu'à stabilisation du syst*

iii bis)- De même les voisins les plus proches seront très légèrement modifiés, la variation dépend la distance $r^k - r^{next}$ qui a été calculée avec la formule de la distance euclidienne

iv)- Le processus recommence depuis le i)- jusqu'à la stabilisation du système. La valeur des modifications diminuera donc avec le temps. La carte se déformera peu à peu pour atteindre un équilibre.

Les paramètres d'apprentissage sont capitaux dans l'apprentissage ainsi la figure 7.5 montre l'évolution des paramètres . Cette influence est fortement perceptible, des coefficients trop faible tendent à dimininuer les mouvements et provoquent une fixation des neurones trop rapides, d'où une mauvaise classification.

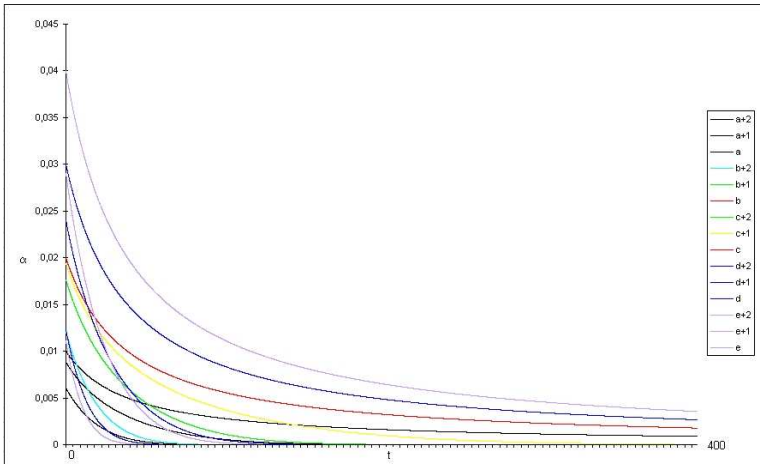


FIG. 7.5 – *Evolutioon des coefficients en fonction des cycles avec diffŕents coeffcients ν et ρ classiques.*

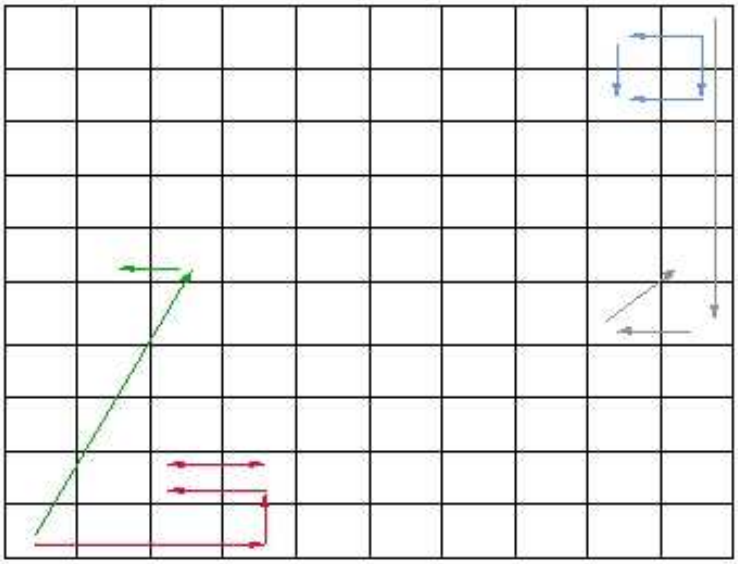


FIG. 7.6 – *Mouvement lors des premiers cycles des neurones les plus proches des structures α et β pour des valeurs d'apprentissage et de diffusion distincte.*

La figure 7.6 montre les différences de mouvement impliqués par des coefficients distincts. Sur l'exemple, le neurone de l'hélice α central a été placée en bas à droite, celui du feuillet en haut à gauche. Selon les coefficients utilisés soit les mouvements sont importants et l'apprentissage semble peu dépendant de l'initialisation, soit ils sont faibles et dépendent fortement de l'intialisation. Ainsi, le neurone β peut ne pas se déplacer et donc se focaliser de maniére "injuste", génant les neurones qui lui sont proches.

k-means La méthode des k-means est proche des cartes auto-organisées de Kohonen [108], mais ne possède aucun processus de diffusion. 5 étapes principales peuvent-être définies :

(1) Il faut définir le nombre b de groupe,

(2) Une observation est associée à chaque groupe de manière aléatoire. et devient le *centre* du groupe.

(3) Le processus dynamique peut alors commencer, il consiste à associer chaque observation de la base de donnée au groupe dont elle est le plus proche, sa distance est minimale avec le centre de ce groupe.

(4) Quand toutes les observations sont assignées à un groupe, Chaque centre est recalculé comme étant la moyenne, le barycentre des observations associées au groupe.

(5) Et le processus recommence depuis l'étape (3) jusqu'à stabilisation du système.

Annexe 2 : Articles

de Brevern AG, Etchebest C, et Hazout, S (2000),"Bayesian probabilistic approach for prediction backbone structures in terms of protein blocks", *Proteins : Structure, Fucntions and Genetics*, 41(3), pp.271-287 [35].

abstract Using an unsupervised cluster analyser, we have identified a local structural alphabet composed of 16 folding patterns of five consecutive C_α ("protein blocks"). The dependence that exists between successive blocks is explicitly taken into account. A Bayesian approach based on the relation protein block-amino acid propensity is used for prediction and leads to a success rate close to 35%. Sharing sequence windows associated with certain blocks into "sequence families" improves the prediction accuracy by 6%. This prediction accuracy exceeds 75% when keeping the first four predicted protein blocks at each site of the protein.

de Brevern AG, et Hazout S (2000),"Hybrid Protein Model (HPM): a method to compact protein 3D-structures information and physicochemical properties", *IEEE - Computer Society : Proceedings of the 7th Symposium on String Processing and Information Retrieval*, pp.49-54 [36].

abstract The transformation of protein 1D-sequence to protein 3D-structure is one of the main difficulties of the structural biology. A structural alphabet had been previously defined from dihedral angles describing the protein backbone as structural information by using an unsupervised classifier. The 16 Protein Blocks (PBs), basis element of the structural alphabet, allows a correct 3D structure approximation [35]. Local prediction had been estimated by a Bayesian approach and shown that sequence information induces strongly the local fold, but stays coarse (prediction rate of 40.7% with one PB, 75.8% with the four most probable PBs).

The Hybrid Protein Model presented in this study learns both sequence and structure of the proteins. The analysis made along the hybrid protein has permitted to appreciate more precisely the spatial location of some types of amino acid residues in the secondary structures and their flanking regions. This study leads to a fuzzy model of dependence between sequence and structure.

de Brevern AG, et Hazout S (2001), "Compacting local protein folds with a Hybrid Protein", *Theoretical Chemistry Accounts*, *sous presse* [37].

abstract The "Hybrid Protein Model" (HPM) is a fuzzy model for compacting local protein structures. It learns a non-redundant database encoded in a previously defined structural alphabet composed of 16 protein blocks (PBs) [35]. The hybrid protein is composed of a series of distributions of the probability of observing the PBs. The training is an iterative unsupervised process that for every fold to be learnt consists of looking for the most similar pattern present in the hybrid protein and modifying it slightly. Finally each position of the hybrid protein corresponds to a set of similar local structures. Superimposing those local structures yields an average root mean square of 3.14 Å. The significant amino acid characteristics related to the local structures are determined. The use of this model is illustrated by finding the most similar folds between two cytochromes P450.

Camproux AC, de Brevern AG, Tufféry P, et Hazout S, "Exploring the use of a structural alphabet for a structural prediction of protein loops", *Theoretical Chemistry Accounts*, *sous presse* [21].

abstract The prediction of loop conformations is one of the challenging problems of homology modeling, due to the large sequence variability associated with these parts of protein structures. In the present study, we introduce a search procedure that evolves in a structural alphabet space deduced from a hidden Markov model to simplify the structural information. It uses a Bayesian criterion to predict, from the amino acid sequence of a loop region, its corresponding word in the structural alphabet space.

Results show, that our approach ranks 30% of the target words with the best score, 50% within the 5 best scores. Interestingly, our approach is also suited to accept or not the prediction performed. This allows to rank 57% of the target words with the best score, 67% within the 5 best scores, accepting 16% of learned words and rejecting 93% of unknown words.

Annexe 3 : Les I-sites

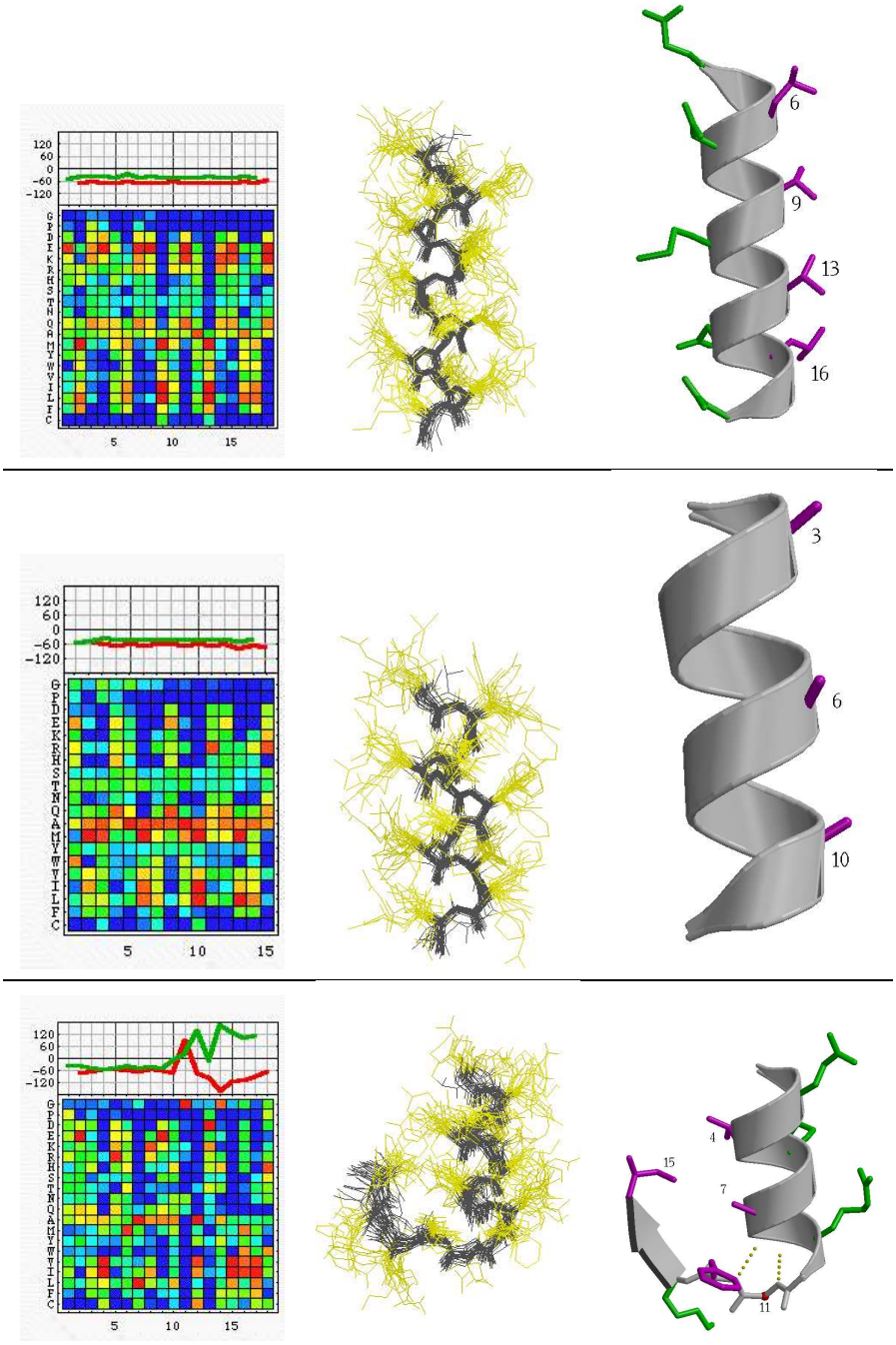


FIG. 7.7 – *I-sites 1 à 3. 1- hélice α amphipatique. 2- hélice α non polaire. 3- extrémité C-terminale d'hélice α Glycine Type 1. (la légende détaillée est sous la figure 7.11).*

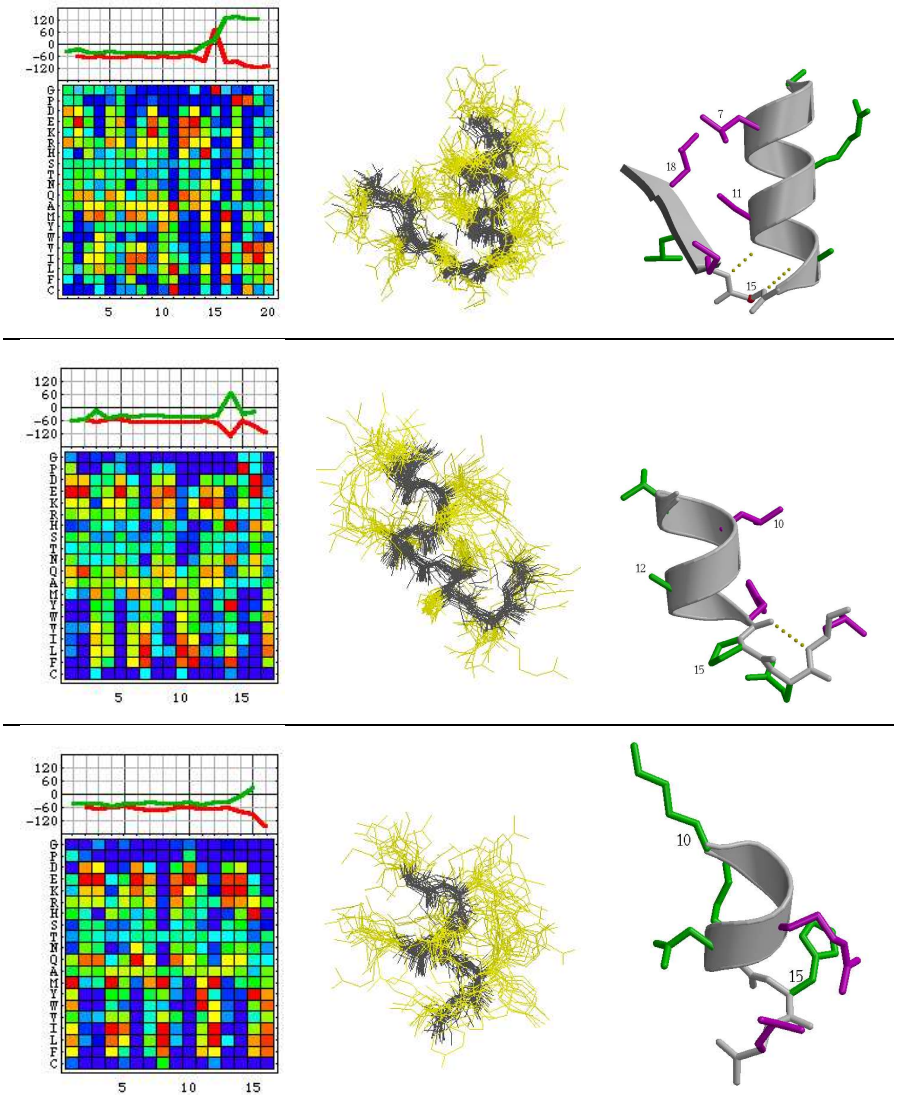


FIG. 7.8 – *I-sites 4 à 6. 4- extrémité C-terminale d'hélice α Glycine Type 2. 5- extrémité C-terminale d'hélice α Proline. 6- hélice α mélée. (la légende détaillée est sous la figure 7.11).*

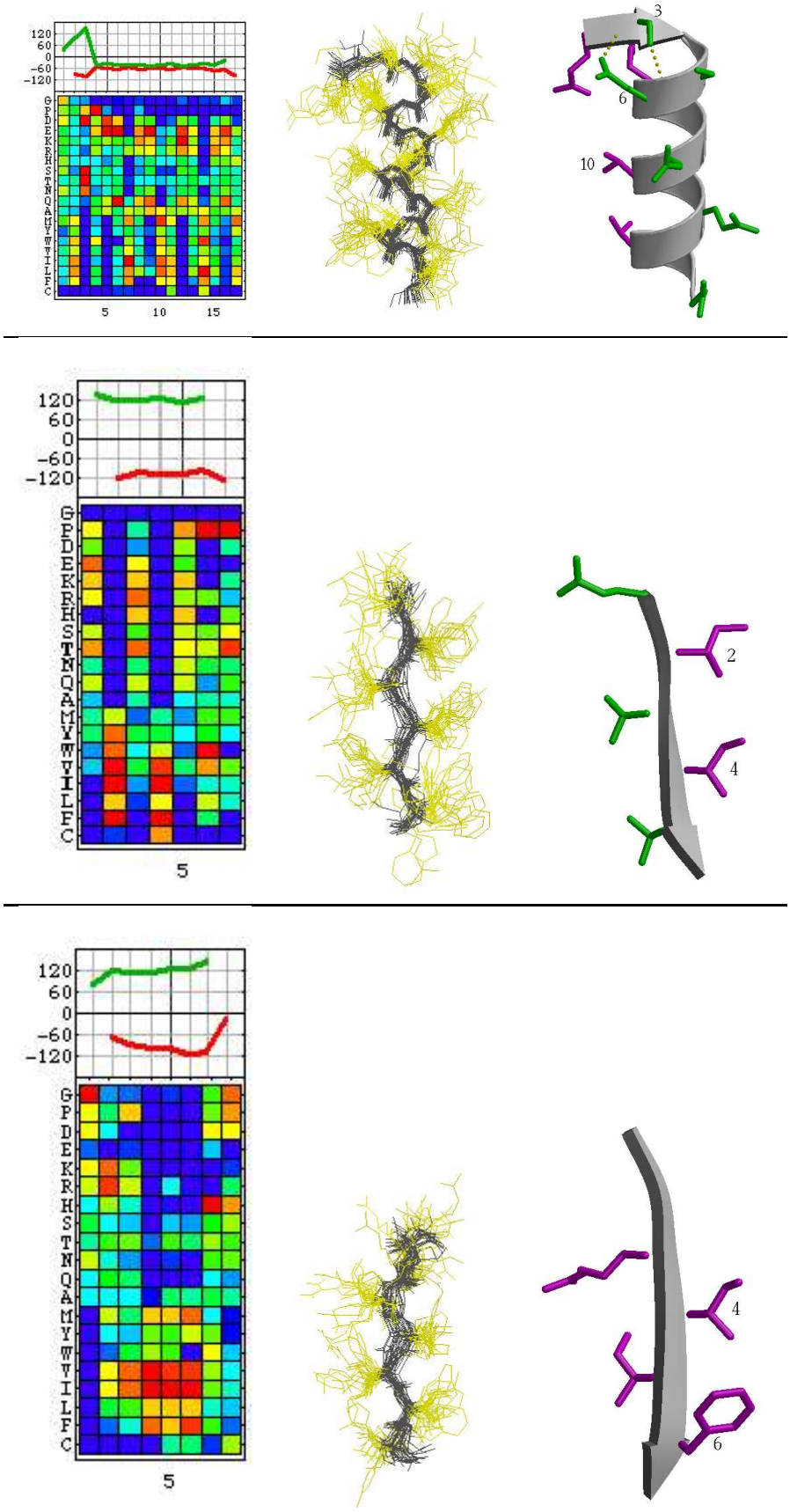


FIG. 7.9 – *I-sites 7 à 9. 7- extrémité N-terminale d'hélice α Sérine. 8- feuillet β amphipatique. 9- feuillet β hydrophobe. (la légende détaillée est sous la figure 7.11).*

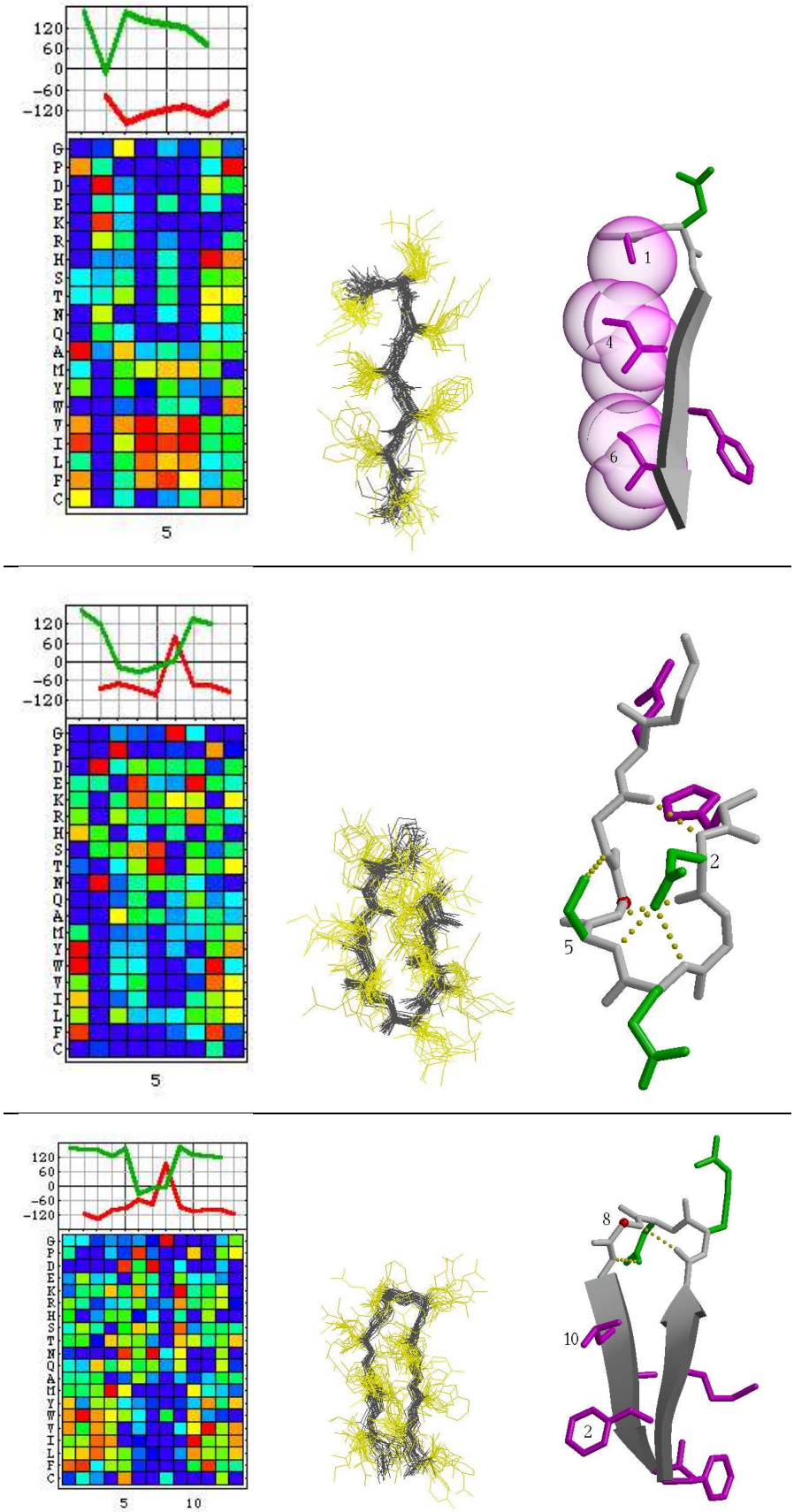


FIG. 7.10 – *I-sites 10 à 12. 10- coude β Aspartate. 11- épingle β Sérine. 12- épingle type I étendu. (la légende détaillée est sous la figure 7.11).*

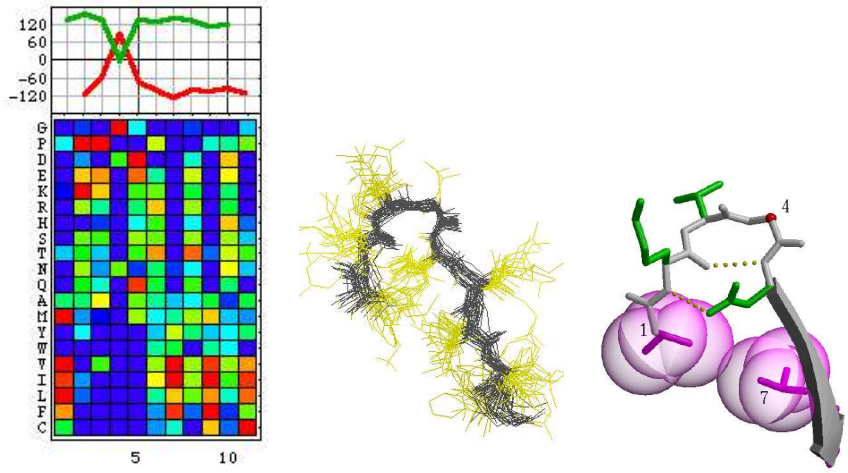


FIG. 7.11 – *I-site 13 coude type II divergeant. underlinelégende: à gauche les angles du squelette polypeptidique (en rouge l'angle ϕ, en bleu l'angle ψ) et en dessous la matrice de contingence du I-site coloré selon l'occurrence de la fréquence de l'acide aminé (en rouge, 3 fois plus que la moyenne de l'acide aminé dans la base de donnée, orange, entre 2 et 3 fois plus, jaune, entre 1 et 2 fois plus, vert, équivalent à la base de donnée, cyan, entre 1 et 2 fois moins, bleu entre 2 et 3 fois moins, bleu marine, plus de 3 fois moins que dans la base de donnée), l'ordre des acides aminés est le suivant en partant du haut: G, P, D, E, K, R, H, S, T, N, Q, A, M, Y, W, V, I, L, R, C; au centre, la superposition des 30 meilleurs représentant du I-site et à droite, une représentation "cartoon" des chaînes latérales (en vert, les polaires, en pourpre, les non-polaires, les glycines sont en rouge).Ces images viennent du site des I-sites http://isites.bio.rpi.edu/Isites/index.html*

Bibliographie

[1] C. André, B. Vincens, J. Boisvieux, and S. Hazout. Mosaic : segmenting multiple aligned dna sequences. *Bioinfomatics*, NA:–, 2001.

[2] R. Aurora and G. Rose. Helix capping. *Protein Science*, 7:21–38, 1998.

[3] D. Barlow and J. Thornton. Helix geometry in proteins. *J Mol Biol*, 201:601–619, 1988.

[4] P. Bates and M. Sternberg. Model building by comparison at casp3 : using expert knowledge and computer automation. *Proteins*, S3:47–54, 1999.

[5] I. Berezovsky, A. Grosberg, and E. Trifonov. Closed loops of nearly standard size: common basic element of protein structure. *FEBS Letters*, 466:283–286, 2000.

[6] F. Bernstein, T. Koetzle, G. Williams, E. Meyer, M. Brice, J. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The protein data bank: a computer-based archival file for macromolecular structures. *J Mol Biol*, 112:535–540, 1977.

[7] J. Bienkowska, R. j. Rogers, and T. Smith. Filtered neighbors threading. *Proteins*, 37:346–359, 1999.

[8] J. Bienkowska, R. j. Rogers, and T. Smith. Protein fold recognition by total alignement probability. *Proteins*, 40:451–462, 2000.

[9] T. Blundell, D. Carney, S. Gardner, F. Hayes, B. Howlin, and T. Hubbard. Knowledge-based protein modelling and design. *Eur. J. Biochem.*, 172:513–520, 1988.

[10] T. Blundell, B. Sibanda, M. Sternberg, and J. Thornton. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature*, 326:347, 1987.

[11] J. Boberg, T. Salakoski, and M. Vihinen. Selection of a representative set of structures from brookhaven protein databank. *Proteins*, 14:264–276, 1992.

[12] T. bottom line for prediction of residue solvent accessibility. Richardson, cj and barlow, dj. *Protein Eng*, 12:1051–1054, 1999.

[13] N. Boutonnet, A. Kajava, and M. Rooman. Structural classification of alphabetabeta and betabetaalpha supersecondary structure. *Proteins*, 30:193–212, 1998.

[14] M. Bower, F. Cohen, and R. Dunbrack Jr. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: A new homology modeling tool. *J. Mol. Biol.*, 267:1268–1282, 1993.

[15] J. Bowie, R. Lűthy, and D. Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253:164–169, 1991.

[16] R. Britten. Precise sequence complementarity between yeast chromosome ends and two classes of just-subtelomeric sequences. *Proc Natl Acad Sci USA*, 26:5906–59120, 1998.

[17] S. Bryant and C. Lawrence. An empirical energy function for threading protein sequence through the folding motif. *Proteins*, 16:92–112, 1993.

[18] C. Bystroff and D. Baker. Blind predictions of local protein structure in casp2 targets using the i-sites library. *Proteins*, supl.1:167–171, 1997.

[19] C. Bystroff and D. Baker. Prediction of local structure in proteins using a library of sequence-structure motif. *J Mol Biol*, 281:565–577, 1998.

[20] C. Bystroff, V. Thorsson, and D. Baker. Hmmstr: a hidden markov model for local sequence-structure correlations in proteins. *J Mol Biol*, 301:173–190, 2000.

[21] A. Camproux, A. de Brevern, P. Tuffery, and S. Hazout. Exploring the use of a structural alphabet for a structural prediction of protein loops. *Theoretical Chemistery Accounts*, page en revision, 2001.

[22] A. Camproux, F. Saunier, G. Chouvet, J. Thalabard, and G. Thomas. A hidden markov model approach to neuron firng patterns. *Biophysical Journal*, 71:2404–2412, 1996.

[23] A. Camproux, P. Tuffery, L. Buffat, C. Andre, J. Boisvieux, and S. Hazout. Analyzing patterns between regular secondary structures using short structural blocks defined by a hidden markov model. *Theoretical Chemistery Accounts*, 101:33–40, 1999.

[24] A. Camproux, P. Tuffery, J. Chevrolat, J. Boisvieux, and S. Hazout. Hidden markov model approach for identifying the modular framework of the protein backbone. *Protein Eng*, 12(12):1063–1073, 1999.

[25] O. Carugo. Predicting residue solvent accessibility from protein sequence by considering the sequence environment. *Protein Eng*, 13:607–609, 2000.

[26] H. Chan. Folding alphabet. *Nature Structural Biology*, 6:994–996, 1999.

[27] J. Chandonia and M. Karplus. The importance of larger data sets for protein secondary structure prediction with neural networks. *Protein Sci.*, 5:768–774, 1996.

[28] J. Chandonia and M. Karplus. New methods for accurate prediction of protein secondary structure. *Proteins*, 35:293–306, 1999.

[29] P. Chou and G. Fasman. Prediction of protein conformation. *Biochemistry*, 13:222–245, 1974.

[30] M. Claessens, E. Cutsen, I. Lasters, and S. Wodak. Modelling the polypeptide backbone with 'spare parts' from known protein structures. *Prot. Eng.*, 4:335, 1989.

[31] N. Colloc'h, C. Etchebest, E. Thoreau, B. Henrissat, and J. Mornon. Comparaison of three algorithms for the assignement of secondary structure in proteins: the advantages of a concenssus assignement. *Protein Eng.*, 6:377–382, 1993.

[32] L. Conte and J. Chothia, C andJanin. The atomic structure of protein-protein recognition sites. *J Mol Biol*, 5:2177–2199, 1999.

[33] J. Cuff and G. Barton. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins*, 34:508–519, 1999.

[34] J. Cuff, M. Clamp, A. Siddiqui, M. Finlay, and G. Barton. Jpred: a consensus secondary structure prediction server. *Proteins*, 14:892–893, 1998.

[35] A. de Brevern, C. Etchebest, and S. Hazout. Bayesian probabilistic approach for prediction backbone structures in terms of protein blocks. *Proteins*, 41(3):271–287, 2000.

[36] A. de Brevern and S. Hazout. Hybrid protein model (hpm): a method to compact protein 3d-structures information and physicochemical properties. *IEEE - Computer Society: Proceedings of the 7th Symposium on String Processing and Information Retrieval*, 1(1):49–57, 2000.

[37] A. de Brevern and S. Hazout. Compacting local protein folds with a hybrid protein. *Theoretical Chemistry Accounts*, NA:–, 2001.

[38] C. Deane and T. Blundell. A novel exhaustive search algorithm for predicting the conformation of polypeptide segments in proteins. *Proteins*, 40:135–144, 2000.

[39] T. Defay and F. Cohen. Evaluation of current techniques for ab initio protein structure prediction. *Proteins*, 23:431–445, 1995.

[40] G. Deléage, C. Blanchet, and C. Geourjon. Protein structure prediction. implications for the biologist. *Biochimie*, 79:681–686, 1997.

[41] P. Derreumaux. Folding a 20 amino acid $\alpha\beta$ peptide with the diffusion process-controled monte-carlo method. *J. Chem. Phys.*, 107:1941–1947, 1997.

[42] V. Di Francesco, J. Garnier, and P. Munson. Improving protein secondary structure prediction with aligned homologous sequences. *Protein Science*, 5:106–113, 1996.

[43] C. Dodge, R. Schneider, and C. Sander. The hssp database of protein structure-sequence alignments and family profiles. *Nucleic Acid Res*, 26:313–315, 1998.

[44] L. Donate, S. Rufino, L. Canard, and T. Blundell. Conformational analysis and clustering of short and medium size loops connecting secondary structures: a database for modeling and prediction. *Protein Science*, 5:2600–2616, 1996.

[45] R. Dunbrack Jr. and M. Karplus. Backbone-dependent rotamer library for proteins: Application to side-chain prediction. *J. Mol. Biol.*, 230:543–574, 1993.

[46] A. Efimov. Super-secondary structures involving triple-strand beta-sheets. *FEBS Lett*, 334:253–256, 1993.

[47] N. Eswar and C. Ramakrishan. Secondary structures without backbone : an analysis of backbone mimicry by ploar side chains in protein structures. *Protein Eng*, 12:447–455, 1999.

[48] P. Fariselli and R. Casadio. A neural network based predictor of residue contacts in proteins. *Protein Eng*, 12:15–21, 1999.

[49] P. Fariselli and R. Casadio. Prediction of the number of residue contacts in proteins. *Ismb. 2000*, 8:146–151, 2000.

[50] P. Fariselli, P. Riccobelli, and R. Casadio. Role of evolutionary information in predicting the disulfide-bonding state of cysteine in proteins. *Proteins*, 36:340–346, 1999.

[51] J. Fetrow. Omega loops: nonregular secondary structures significant in protein function and stability. *FASEB J*, 9:708–717, 1995.

[52] J. Fetrow and G. Berg. Using information theory to discover side chain rotamer classes: analysis of the effects of local backbone structure. *Pacific Symposium*, 1:–, 1998.

[53] J. Fetrow, M. Palumbo, and G. Berg. Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary structure classification scheme. *Proteins*, 27:249–271, 1997.

[54] A. Finkelstein and B. Reva. A search for the most stable folds of protein chains. *Nature*, 351:497–499, 1991.

[55] A. Finkelstein and B. Reva. Search for the most stable folds of protein chains: I. application of a self-consistent molecular field to a problem of protein three-dimensional structure prediction. *Protein Eng*, 9:387–397, 1996.

[56] A. Fiser, N. Dosztányi, and S. I. The role of long-range interactions in defining the secondary structure of proteins is overestimated. *Comput. Apll. Biosci*, 13:297–301, 1997.

[57] A. Fiser and I. Simon. Predicting the oxidation state of cysteines by multiple sequence alignment. *Bioinformatics*, 16:251–256, 2000.

[58] D. Frishman and P. Argos. Knowledge-based protein secondary structure. *Proteins*, 23:566–579, 1995.

[59] D. Frishman and P. Argos. The future of protein secondary structure accuracy. *Fold. Des.*, 2:159–162, 1997.

[60] D. Frishman and P. Argos. Seventy-five percent accuracy in protein secondary structure prediction. *Proteins*, 27:97–120, 1997.

[61] J. Garnier, J. Gibrat, and B. Robson. Gor method for predicting protein secondary structure from amino acid sequence. *Methods. Enzymol*, 266:540–553, 1996.

[62] J. Garnier, D. Osguthorpe, and B. Robson. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol*, 120:97–120, 1978.

[63] V. Geetha and P. Munson. Linkers of secondary structures in proteins. *Protein Science*, 6:2538–2547, 1997.

[64] C. Geourjon and G. Deléage. Sopma: significant improvements in protein secondary structure prediction by concensus prediction from multiple alignements. *CABIOS*, 6:681–684, 1995.

[65] J.-F. Gibrat, J. Garnier, and B. Robson. Further developments of protein secondary structure prediction using information theory. *J Mol Biol*, 198:425–443, 1987.

[66] J.-F. Gibrat, B. Robson, and J. Garnier. Inluence of the local amino acid sequence upon the zones of the torsional angles ϕ and ψ adopted by residues in proteins. *Biochemistery*, 30:1578–1586, 1991.

[67] J. Gorodkin, O. Lund, C. Andersen, and S. Brunak. Using sequence motifs for enhanced neural network prediction of protein distance constraints. *Proceedings of the seventh international conference for molecular biology (ISMB'99)*, 1:95–105, 1999.

[68] C. Gourgeon and G. Deleage. Sopm: a self-optimized method for protein secondary structure prediction. *Prot. Eng.*, 7:157–164, 1994.

[69] S. Govindarajan and R. Goldstein. Why are some proteins structures so common? *Proc. Natl. Acad. Sci. USA*, 93:3341–3345, 1996.

[70] S. Govindarajan, R. Recabarren, and R. Goldstein. Estimating the total number of protein folds. *Proteins*, 35:408–414, 1999.

[71] M. Gribskov, A. McLachlan, and D. Eisenberg. Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A*, 84:4355–4358, 1987.

[72] Y. Guermeur, C. Geourjon, P. Gallinari, and G. Deléage. Improved performance in protein secondary structure prediction by inhomogeneous score combination. *Bioinformatics*, 15:413–421, 1999.

[73] C. Hadley and D. Jones. A systematic comparison of protein structure classications scop, cath and fssp. *Structure*, 7:1099–1112, 1999.

[74] K. Han and D. Baker. Recurring local sequence motifs in proteins. *J Mol Biol*, 241:176–187, 1995.

[75] K. Han and D. Baker. Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proc. Natl. Acad. Sci. USA*, 93:5814–5818, 1996.

[76] K. Han, C. Bystroff, and D. Baker. Three dimensional structures and contexts associated with recurrent amino acid sequence patterns. *Protein Science*, 6:1587–1590, 1997.

[77] J. Hanke and J. Reich. Kohonen map as a visualization tool for the analysis of protein sequences: multiple alignments, domains and segments of secondary structures. *Comput. Appl. Biosci.*, 12:447–454, 1996.

[78] J. Hanke and J. Reich. Self-organizing hierarchic networks for pattern recognition in protein sequence. *Protein Science*, 5:72–82, 1996.

[79] J. Hartigan and M. Wong. k-means. *Applied Statistics*, 28(1):100–115, 1979.

[80] C. Haseman, R. Kurumbail, N. Boddupalli, J. Peterson, and N. Deisenhofer. Structure and function of cytochromes p450: a comparative analysis of three crystal structures. *Structure*, 2:41–62, 1995.

[81] C. Hasemann, K. Ravichandran, J. Peterson, and J. Deisenhofer. Crystal structure and refinement of cytochrome p450terp at 2.3 a resolution. *J Mol Biol*, 4:1169–1185, 1994.

[82] S. Hayward and J. Collins. Limits on α-helix prediction with neural networks models. *Proteins*, 14:372–381, 1992.

[83] U. Hobohm and C. Sander. Enlarged representative set of protein structures. *Prot Sci*, 3:522–524, 1994.

[84] U. Hobohm, F. Scharf, R. Schneider, and C. Sander. Selection of a representative set of structures from the brookhaven protein databank. *Prot Sci*, 1:409–417, 1992.

[85] L. Holley and M. Karplus. Protein secondary structure prediction with a neural network. *Proc. Natl. Acad. Sci.*, 86:152–156, 1989.

[86] L. Holm and C. Sander. The fssp database of structurally aligned protein fold families. *Nucl. Acid. Res.*, 22:3600–3609, 1994.

[87] L. Holm and C. Sander. Dali/fssp classification of three-dimensional protein folds. *Nucl Acids Res*, 25:231–234, 1997.

[88] E. Huang, R. Samudrala, and J. Ponder. Ab initio fold prediction of small helical proteins using distance geometry and knowledge-based scoring functions. *J Mol Biol*, 290:267–281, 1999.

[89] T. Hubbard, A. Murzin, S. Brenner, and C. Chotia. Scop : a structural classification of proteins database. *Nucleic Acids Research*, 25:236–239, 1997.

[90] E. Hutchinson and J. Thornton. A revised set of potentials for β-turn formation in protein. *Prot Sci*, 3:2207–2216, 1994.

[91] J. Janin. Surface area of globular proteins. *J. Mol. Biol.*, 105:13, 1976.

[92] J. Janin. Wet and dry interfaces: the role of solvent in protein-protein and protein-dna recognition. *Structure Fold Des*, 7:R277–279, 1999.

[93] L. Jaroszewski, L. Rychlewski, B. Zhang, and A. Godzik. Fold prediction by a hierarchy of sequence, threading, and modeling methods. *Prot Sci*, 7:1431–40, 1998.

[94] P. Jean, J. Pothier, P. Dansette, D. Mansuy, and A. Viari. Automated multiple analysis of protein structures: application to homology modeling of cytochromes p450. *Proteins*, 28:388–404, 1997.

[95] D. Jones, W. Taylor, and J. Thornton. A new approach to protein fold recognition. *Nature*, 358:86–89, 1992.

[96] D. Jones and J. Thornton. Protein fold recognition. *Journal of Computer-Aided Molecular Design*, 7:439–456, 1993.

[97] D. Jones, M. Tress, K. Bryson, and C. Hadley. Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure. *Proteins*, S3:104–111, 1999.

[98] S. Jones, M. Stewart, A. Michie, M. Swindells, C. Orengo, and J. Thornton. Domain assignment for protein structures using a concensus approach ; characterization and analysis. *Prot Sci*, 7:233–242, 1998.

[99] T. Jones and S. Thirup. Using know substructures in protein model building and crystallography. *EMBO J*, 5:819–822, 1986.

[100] W. Kabsh and C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features. *Biopolymers*, 22:2577, 1983.

[101] N. Kannan and S. Vishveshwara. Identification of side-chain clusters in protein structures by a graph spectral method. *J. Mol. Biol*, 292:441–464, 1999.

[102] M. Karpen, P. de Haseth, and K. Neet. Comparing short protein substructures by a method based on backbone torsion angles. *Proteins*, 6:155–167, 1989.

[103] M. Karpen, P. de Haseth, and K. Neet. Differences in the amino acid distributions of 3_10-helices and α-helices. *Protein Science*, 1:1333–1342, 1992.

[104] T. Kawabata and J. Doi. Improvement of protein secondary structure prediction using binary word encoding. *Proteins*, 27:36–46, 1997.

[105] R. King and M. Sternberg. Identification and application of the concepts important fro accurate and reliable protein secondary structure prediction. *Protein Science*, 5:2298–2310, 1996.

[106] I. Koch, T. Lengauer, and E. Wanke. An algoritm for finding maximal common subtopologies in a set of protein structures. *Journal of Computational Biology*, 3:289–306, 1996.

[107] T. Kohonen. Learning vector quantization. *Neural Networks*, 1(suppl. 1):303, 1989.

[108] T. Kohonen. Statistical pattern recognition revisited. *Advanced Neural Computers,**R. Eckmiller (editor)**,Elesevier Science publisher (Holland)*, pages –, 1990.

[109] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, Berlin, Germany, 1997.

[110] A. Kolinski, P. Rotkiewicz, B. Ilkowski, and J. Skolnick. A method for the improvement of threading-based protein models. *Proteins*, 37:592–610, 1999.

[111] J. Kraulis. Molscript: A program to produce both detailed and schematic plots of protein structures. *J Appl Cryst*, 24:946–950, 1991.

[112] S. Kullback and R. Leibler. On information and sufficiency. *Ann Math Stat*, 22:79–86, 1951.

[113] S. Kumar and M. Bansal. Geometrical and sequence characterisitics of α-helics in globular proteins. *Biophysical Journal*, 78:1935–1944, 1998.

[114] J.-M. Kwasigroch, J. Chomilier, and J.-P. Mornon. A global taxonomy of loops in globular proteins. *J Mol Biol*, 259:855–872, 1996.

[115] J. Kyte and R. Doolittle. A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, 157(1):105–132, 1982.

[116] G. Labesse, N. Colloc'h, J. Pothier, and J.-P. Mornon. P-sea: a new efficient assignement of secondary structure from cα. *Comput. Appl. Biosci.*, 13:291–295, 1997.

[117] R. Lathrop, R. Rogers Jr., T. Smith, and J. White. A bayes-optimal sequence-structure theory that unifies protein sequence-structure recognition and alignement,. *Bulletin of Mathematical Biology*, 60:1039–1071, 1998.

[118] B. Lee and F. Richards. The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.*, 55:379–392, 1971.

[119] C. Levinthal. Molecular model-building by computer. *Sci Am*, 214:42–52, 1966.

[120] M. Levitt. Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.*, 226:507–533, 1992.

[121] R. Lűthy, D. MCLachlan, and D. Eisenberg. Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins*, 10:229–239, 1991.

[122] M. Liebman, C. Venanzi, and H. Weinstein. Structural analysis of carboxypeptidase a and its complexes with inhibitors as a basis for modeling enzyme recognition and specificity. *Biopolymers*, 24:1721–1758, 1985.

[123] L. Lo Conte, B. Ailey, T. Hubbard, S. Brenner, A. Murzin, and C. Chotia. Scop : a structural classification of proteins database. *Nucleic Acid Research*, 28:257–259, 2000.

[124] O. Lund, K. Frimand, J. Gorodkin, H. Bohr, J. Hansen, and S. Brunak. Protein distance constraints predicted by neural networks and probability density functions. *Protein Engineering*, 10:1241–1248, 1997.

[125] M. MacArthur and J. Thornton. Deviations from planarity of peptide bond in peptides and proteins. *J Mol Biol*, 264:1180–1195, 1996.

[126] T. Madej, J.-F. Gibrat, and S. Bryant. Threading a database of protein cores. *Proteins*, 23:356–369, 1995.

[127] A. Michie, C. Orengo, and J. Thornton. Analysis of domain structural class using an automated class assignement protocol. *J Mol Biol*, 262:168–85, 1996.

[128] Mucchielli-Giorgi. Thèse. *Paris 7*, 5:70–122, 1999.

[129] M. Mucchielli-Giorgi, S. Hazout, and T. P. Predacc: prediction of solvent accessibility. *Bioinformatics*, 15:176–177, 1999.

[130] M. Mucchielli-Giorgi, T. P, and S. Hazout. Prediction of solvent accessibility of amino acid residues: critical aspects. *Theoretical chemistry accounts*, 101:186–193, 1999.

[131] A. Murzin. Structural classification of proteins: new superfamilies. *Current opinion in Structural Biology*, 6:386–394, 1996.

[132] A. Murzin, S. Brenner, T. Hubbard, and C. Chotia. Scop : a structural classification of proteins database for the investigation of sequences and structures. *JMB*, 247:526–540, 1995.

[133] S. Muskal and S. Kim. Predicting protein secondary structure content. a tandem neural network approach. *J Mol Biol*, 225:713–727, 1992.

[134] K. Nadassy, S. Wodak, and J. Janin. Structural features of protein-nucleic acid recognition sites. *Biochemistry*, 38:1999–2017, 1999.

[135] O. Olmea and A. Valencia. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold Des*, 2:S25–32, 1997.

[136] C. Orengo. Classification of protein folds. *Current opinion in Structural Biology*, 4:429–44, 1994.

[137] C. Orengo, J. Bray, T. Hubbard, L. LoConte, and I. Sillitoe. Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. *Proteins*, suppl.3:149–170, 1999.

[138] C. Orengo, S. Jones, and J. Thornton. Protein superfamilies and domain superfolds. *Nature*, 372:631–634, 1994.

[139] C. Orengo, A. Michie, J. Jones, M. Swinells, and J. Thornton. Cath- a hierarchic classification of protein domain structures. *Structure*, 5:1093–1098, 1997.

[140] C. Orengo, F. Peral, J. Bray, A. Todd, A. Martin, L. Lo Conte, and J. Thornton. The cath database provides insights into protein structure/function relationships. *Nucleic Acids Research*, 27:275–279, 1999.

[141] D. Osguthorpe. ab initio protein folding. *Current Opinion in Structural Biology*, 10:146–152, 2000.

[142] M. Ouali and K. RD. Cascaded multiple classifiers for secondary structure prediction. *Protein Science*, 9:1136–1176, 2000.

[143] A. Panchenko, A. Marchler-Bauer, and S. Bryant. Threading with explicit models for evolutionary conservation of structure and sequence. *Proteins*, Suppl 3:133–140, 1999.

[144] T. Petersen, C. Lundegaard, M. Nielsen, H. Bohr, J. Bohr, S. Brunak, G. Gippert, and O. Lund. Prediction of protein secondary structure at 80% accuracy. *Proteins*, 41:17–20, 2000.

[145] L. Presta and G. Rose. Helix signals in proteins. *Science*, 240:1632–1641, 1988.

[146] S. Prestelski, A. Williams Jr., and M. Liebman. Generation of a substructure library for the description and classification of protein secondary structure. i. overview of the methods ans results. *Proteins*, 14:430–439, 1992.

[147] S. Prestelski, A. Williams Jr., and M. Liebman. Generation of a substructure library for the description and classification of protein secondary structure. ii. application to spectra-structure correlations in fourrier transform infrared spectroscopy. *Proteins*, 14:440–450, 1992.

[148] N. Qian and T. Sejnowski. Prediction of secondary structure of globular proteins using a neural network. *J. Mol. Biol.*, 202:865–884, 1988.

[149] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77:257–285, 1989.

[150] K. Rajashankar and S. Ramakumar. Pi-turns in proteins and peptides: classification, conformation, occurence, hydratation and sequence. *Protein Sci*, 5:932–946, 1996.

[151] K. Ravichandran, S. Boddupalli, C. Hasemann, J. Peterson, and J. Deisenhofer. Crystal structure of hemoprotein domain of p450bm-3, a prototype for microsomal p450's. *Science*, 6:731–736, 1993.

[152] B. Reva, A. Finkelstein, and J. Skolnick. What is the probability of a chance prediction of a protein structure with an rmsd of 6 Å? *Fold Des*, 3:141–147, 1998.

[153] B. Reva, J. Skolnick, and A. Finkelstein. Averaging interaction energies over homologs improves protein recognition in gapless threading. *Proteins*, 35:353–359, 1999.

[154] J. Richardson and D. Richardson. Amino acid preferences for specific locations at the end of α helices. *Science*, 240:1648–1652, 1988.

[155] T. Richmond. Solvent accessible surface area and excluded volume in proteins. *J. Mol. Biol.*, 178:63–89, 1984.

[156] F. Ridchards and C. Kundrot. Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure,. *Proteins*, 3:71–84, 1988.

[157] C. Rohl and A. Doig. Models for the 3.10-helix/coil, π-helix-coil and α-helix/3.10-helix/coil transitions in isolated peptides. *Protein Science*, 5:1689–1696, 1996.

[158] J. Rooman, MJ, S. Wodak, and J. Thronton. Amino acid sequence templates derived from recurrent turn motifs in proteins: critical evaluation of their predictive power. *Protein Eng*, 3:23–27, 1989.

[159] M. Rooman, J.-P. Kocher, and S. Wodak. prediction of protein backbone conformation based on seven structure assignements. influence of local intercations. *J Mol Biol*, 221:961–979, 1991.

[160] M. Rooman, J.-P. Kocher, and S. Wodak. Extracting information on folding from amino acid sequence: accurate predictions for protein regions with preferred conformation in the absence of tertiary interactions. *Biochemistery*, 27:10226–10238, 1992.

[161] M. Rooman, J.-P. Kocher, and S. Wodak. Extracting information on folding from amino acid sequence: concensus regions with preferres conformation in homologous proteins. *Biochemistery*, 27:10239–10249, 1992.

[162] M. Rooman, J. Rodriguez, and S. Wodak. Automatic definition of reccurent local structure motifs in proteins. *J Mol Biol*, 213:327–336, 1990.

[163] M. Rooman, J. Rodriguez, and S. Wodak. Relations between protein sequece and structure and their significance. *J Mol Biol*, 213:337–350, 1990.

[164] M. Rooman and S. Wodak. Identification of predictive sequence motifs limited by protein structure data base size. *Nature*, 335:45–49, 1988.

[165] B. Rost. Phd : predicting one-dimensional protein structure by profile-based neural networks. *Methods. Enzymol*, 232:525–539, 1996.

[166] B. Rost and C. Sander. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol*, 232:584–599, 1993.

[167] B. Rost, C. Sander, and R. Schneider. Phd - an automatic mail server for protein secondary structure prediction. *CABIOS*, 10:53–60, 1994.

[168] A. Salamov and V. Solovyev. Protein secondary structure prediction using local alignenments. *J Mol Biol*, 268:31–36, 1997.

[169] A. Sali and T. Blundell. Compartative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, 234:779–815, 1993.

[170] R. Samudrala and J. Moult. A graph-theoric algorithm for comparative modeling of protein structure. *J Mol Biol*, 279:287–302, 1998.

[171] R. Samudrala, Y. Xia, E. Huanh, and L. M. Ab initio protein structure prediction using a combined hierarchical approach. *Proteins*, 3 suppl.:194–198, 1999.

[172] C. Sander and R. Schneider. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9:56–68, 1991.

[173] F. Sasagawa and K. Tajima. Prediction of protein secondary structures by a neural network. *Comput. Appl. Biosci*, 9:147–152, 1993.

[174] R. Schneider, A. de Daruvar, and C. Sander. The hssp database of protein structure-sequence alignements. *Nucl Acids Res*, 25:226–230, 1997.

[175] J. Schuchhardt, G. Schneider, J. Reichelt, D. Schomburg, and P. Wrede. Local structural motifs of protein backbones are classified by self-organizing neural networks. *Protein Eng*, 9(10):833–842, 1996.

[176] J. Segrest, H. De Loof, J. Dohlman, C. Brouillette, and G. Anantharamaiah. Amphipathic helix motif: classes and properties. *Proteins*, 8:103–117, 1990.

[177] E. Shakhanovich. Modeling protein folding: the beauty of power and simplicity. *Fold. Des.*, 1:50–54, 1996.

[178] K. Simons, R. Bonneau, I. Ruczinski, and D. Baker. Ab initio protein structure prediction of casp 3 targets using rosetta. *Proteins*, 34(suppl. 3):171–176, 1999.

[179] K. Simons, C. Kooperberg, E. Huang, and B. D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J Mol Biol*, 268:209–225, 1997.

[180] K. Simons, I. Ruczinski, C. Kooperberg, B. Fox, C. Bystroff, and D. Baker. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins*, 34:82–95, 1999.

[181] H. Sklenar, C. Etchebest, and R. Lavery. Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis. *Proteins*, 6:46–60, 1989.

[182] V. Solovyev and A. Salamov. Predictiong α-helix and β-strand segments of globular proteins. *CABIOS*, 10:661–669, 1994.

[183] R. Srinivasan and G. Rose. Linus : a hierarchic procedure to predict the fold of a protein. *Proteins*, 22:81–99, 1997.

[184] R. Srinivasan and G. Rose. A physical basis for protein secondary structure. *PNAS*, 96:14258–14263, 1999.

[185] M. Sternberg, P. Bates, L. Kelley, and M. MacCalllum. Progress in protein structure predicition : assessment of casp 3. *Current opinion in Structural Biology*, 9:368–373, 1999.

[186] M. Sternberg and S. Islam. Local protein sequence similarity does not imply a structural relationship. *Protein Eng*, 4:125–131, 1990.

[187] P. Storloz, A. Lapedes, and Y. Xia. Predicting protein secondary structure using neural net and statistical methods. *J. Mol. Biol.*, 225:363–377, 1992.

[188] J. Sun and A. Doig. Addition of side-chain interactions to 3_{10}-helix/coil and α-helix/3_{10}-helix/coil theory. *Protein Science*, 7:2374–2383, 1998.

[189] X. Sun, X. Rao, L. Peng, and D. Xu. Prediction of protein supersecondary structures based on the artificial neural network method. *Protein Eng*, 10:763–769, 1997.

[190] W. Taylor. The classification of amino acid conservation. *J Theor Biol*, 119:205–218, 1986.

[191] W. Taylor. Protein structural domain identification. *Protein Eng*, 12:203–216, 1999.

[192] M. Thompson and R. Goldstein. Predicting solvent accessibility: higher accuracy using bayesian statistics and optimized residue substitution classes. *Proteins*, 25:38–47, 1996.

[193] M. Thompson and R. Goldstein. Predicting protein secondary structure with probabilistic schemata of evolutionarily derived information. *Protein Sci*, 6:1963–1975, 1997.

[194] P. Tuffery. Xmmol : an x11 and motif program for macromolecular visualization and modeling. *J Mol Graphics*, 72:67–72, 1995.

[195] P. Tufféry, C. Etchebest, and S. Hazout. Prediction of protein side chains conformations : a study on the influence of backbone accuracy on conformation stability in the rotamer space. *Prot. Eng.*, 10:361–372, 1997.

[196] P. Tufféry, C. Etchebest, S. Hazout, and R. Lavery. A new approach to the rapid determination of protein side chain conformations. *J. Biomol. Struct. Dynam.*, 6:1267–1289, 1991.

[197] P. Tufféry, C. Etchebest, S. Hazout, and R. Lavery. A critical comparison of search algorithms applied to the optimisation of protein side-chain conformations. *J. Comp. Chem.*, 14:790–798, 1993.

[198] R. Unger, D. Harel, W. S, and S. JL. A 3d building blocks approach to analyzing and predicting structure of proteins. *Proteins*, 5:355–373, 1989.

[199] R. Unger, D. Harel, W. S, and S. JL. Analysis of dihedral angles distribution: the doublets distribution determines polypeptides conformation. *Biopolymers*, 30:499–508, 1990.

[200] R. Unger and S. JL. The importance of short structural motifs in protein structure analysis. *J Comput Aid Mol Des*, 7:457–472, 1993.

[201] L. Wernish, M. Hunting, and S. Wodak. Identification of structural domains in proteins by a graph heuristic. *Proteins*, 36:338–352, 1999.

[202] C. Wilmot and J. Thornton. Analysis and prediction of the different types of beta-turn in proteins. *J Mol Biol*, 203:221–232, 1988.

[203] R. Wintjens, M. Rooman, and S. Wodak. Automatic classification and analysis of alpha alpha-turn motifs in proteins. *J. Mol. Biol*, 255:235–253, 1996.

[204] J. Wodjick, J.-P. Mornon, and J. Chomilier. New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. *J Mol Biol*, 289:1469–90, 1999.

[205] Y. Xu, D. Xu, and E. Uberbacher. An efficient computational method for globally optimal threading. *J Comput Biol*, 5:597–614, 1998.

[206] Q. Yi, C. Bystroff, P. Rajagopal, R. Klevit, and D. Baker. Prediction and structural characterization of an independant folding substructure in the src sh3 domain. *J Mol Biol*, 283:293–300, 1998.

[207] A. Zamyatin. Protein volume in solution. *Prog Biophys Mol Biol*, 24:107–123, 1972.

[208] X. Zhang, J. Fetrow, W. Rennie, D. Waltz, and G. Berg. Automatic derivation of substructures yoelds novel structure building blocks in globular proteins. *ISMB 93*, 1:438–446, 1993.

[209] X. Zhang, J. Fetrow, W. Rennie, D. Waltz, and G. Berg. Design of an auto-associative neural network with hidden layer activation that were used to reclassify local protein structures. *Techniques in protein chemistery V*, 1:397–404, 1994.

[210] K. Zimmermann. When awaiting 'bio' champollion; dynamic programming regularization of the protein secondary structure predictions. *Protein Eng*, 7:1197–1202, 1994.

[211] K. Zimmermann and J. Gibrat. In unison : regularization of protein secondary structure predictions that makes use of multiple alignements. *Prot. Eng.*, 11:865–865, 1998.